DB non solo diritto bancario

leggi l'articolo nel sito www.dirittobancario.it

ATTUALITÀ

Dati sintetici nelle operazioni M&A e nell'acquisto di soluzioni AI

Due diligence, allocazione del rischio e controlli per banche e assicurazioni

8 Settembre 2026

Alessandro Ferrari, Partner, Head of Technology Sector, DLA Piper

Alessandro Ferrari, Partner, Head of Technology Sector, DLA Piper

Alessandro Ferrari

Alessandro Ferrari è Partner del dipartimento Intellectual Property & Technology e dirige il Sector Technology in Italia e si occupa principalmente di diritto applicato alla tecnologia, assistendo i clienti in questioni transactional, advisory and IT litigation. Ha esperienza nella redazione e negoziazione di accordi strategici e cross-border di outsourcing di processi aziendali e di altri contratti commerciali, IT e relativi alla proprietà intellettuale in diversi settori e in diverse geografie. Ha inoltre esperienza nella consulenza su tutti gli aspetti del processo di sourcing/approvvigionamento, compreso lo sviluppo della struttura dell'accordo, la negoziazione e l'assistenza ai clienti nell'implementazione e nelle strategie di integrazione, nella governance e nei performance management regimes.

Studio Legale Tributario Associato

DLA Piper

DLA PIPER

1. Introduzione

I dati sintetici sono record artificiali generati mediante regole, simulazioni o modelli statistici e di machine learning per riprodurre determinate proprietà di dati reali o scenari definiti. Possono essere interamente sintetici oppure sostituire o integrare solo parte di un dataset; soprattutto, sono creati per uno o più compiti specifici. **La loro utilità è quindi dipendente dall'uso previsto**, non una qualità assoluta.

Per banche e assicurazioni l'interesse è concreto: eventi rari possono arricchire dati antifrode o antiriciclaggio; ambienti di test possono essere alimentati senza esporre direttamente archivi di clienti; dati sui sinistri possono essere condivisi con sviluppatori e fornitori in modo più controllato.

Quando questi dati sostengono il prodotto di una target o una soluzione AI acquistata da un intermediario, diventano però parte dell'oggetto economico dell'operazione. Il compratore deve poter rispondere a una domanda semplice: **se domani venisse meno il venditore, il dataset originario o il fornitore che gestisce il generatore - da intendersi il sistema o processo tecnico che produce i dati sintetici, a partire da dati sorgente, regole e parametri - il risultato sarebbe riproducibile, utilizzabile e difendibile verso le funzioni di controllo e l'autorità?**

2. Che cosa si compra davvero

Un file contenente milioni di righe sintetiche è soltanto una fotografia. Il valore risiede in un insieme più ampio: **dataset, generatore, genealogia, parametri, codice e dipendenze, diritti sui dati sorgente, evidenze di validazione e capacità di aggiornamento**. Acquistare solo l'output equivale spesso ad acquistare una versione destinata a invecchiare.

La prima verifica riguarda la funzione dichiarata. Un dataset può essere adeguato per testare la tenuta di un'interfaccia e inadeguato per addestrare un sistema di credit scoring. Può preservare bene le distribuzioni medie e perdere le code, proprio dove si collocano frodi, default e sinistri più gravi. Può aumentare la rappresentazione di un gruppo e introdurre correlazioni artificiali con CAP, professione o fascia reddituale.

Ne deriva una prima regola: **il dataset va valutato rispetto a ciascun uso contrattualmente identificato**. Le espressioni "*high quality*", "*privacy safe*" o "*representative*" non dovrebbero costituire criteri di accettazione finché non sono collegate a popolazione, periodo, variabili, soglie, metriche e decisioni interessate.

3. "Sintetico" non significa automaticamente "anonimo"

"Dato sintetico" descrive un metodo di produzione, non uno status giuridico. Se il generatore è addestrato su dati personali, quella fase resta soggetta al GDPR e richiede finalità, base giuridica, minimizzazione, tempi di conservazione, sicurezza e, quando necessario, una valutazione d'impatto. Anche l'output può conservare informazioni estraibili o consentire inferenze e collegamenti con fonti esterne.

Le **Linee guida 02/2026 dell'EDPB sull'anonimizzazione**, adottate il 7 luglio 2026 e, alla data del presente contributo, ancora in consultazione pubblica, propongono una verifica basata su tre criteri: assenza di isolamento del record, di collegamento e di inferenza. L'anonimato può inoltre variare secondo la prospettiva del soggetto che riceve o controlla i dati e secondo i mezzi ragionevolmente utilizzabili, anche tramite terzi.[1]

In una due diligence non basta quindi un certificato in cui il venditore qualifica il dataset come "anonimo". Occorre acquisire **il rapporto di anonimizzazione**, l'attore e il contesto rispetto ai quali è stato svolto, le informazioni ausiliarie considerate, l'accesso al generatore, le prove di *record linkage* (cioè di collegamento tra registrazioni, volto a ricondurle allo stesso soggetto), l'estrazione e la data dell'ultima rivalutazione. Per i modelli AI addestrati con dati personali, l'EDPB richiede già una valutazione caso per caso della probabilità di estrarre dati di addestramento o di ottenerli mediante interrogazioni.[2]

La prova va ripetuta quando cambiano destinatari, ambiente di accesso, fonti esterne o capacità tecniche. **L'anonimato è una conclusione documentata e contestuale, non una garanzia intrinseca del prodotto.**

4. Qualità, bias e rischio prudenziale

Per i sistemi ad alto rischio, l'AI Act collega la governance dei dati all'origine e alla finalità iniziale della raccolta, alle operazioni di preparazione, lacune, rappresentatività, rilevazione e mitigazione dei bias. Tra gli usi dell'Allegato III rientrano la valutazione del merito creditizio delle persone fisiche e il risk assessment e pricing nelle assicurazioni vita e malattia. A seguito del Regolamento (UE) 2026/1744, il relativo regime è destinato ad applicarsi dal 2 dicembre 2027; la data differita non riduce l'esigenza di costruire oggi la documentazione che dovrà accompagnare un asset acquistato. Il nuovo articolo 4-bis considera espressamente dati sintetici o anonimizzati tra le alternative da esaminare prima del trattamento eccezionale di categorie particolari per correggere bias.[3]

Le aspettative settoriali sono già più vicine. Per i sistemi coperti dalla Opinion del 2025, EIOPA richiede alle imprese di assicurazione di applicare ai dati esterni gli stessi standard di qualità dei dati interni e di conservare evidenze su dati di training, test e metodologie, così da consentire riproducibilità e tracciabilità. L'impresa resta responsabile anche quando il sistema proviene da terzi. La vigilanza BCE richiama, per le banche, rappresentatività, tracciabilità e bias anche come questioni prudenziali, quando producano sottostima o errata determinazione del rischio.[4]

Il test operativo deve separare almeno ***fidelity* (fedeltà statistica rispetto ai dati sorgente), *utility* (utilità per l'impiego previsto), *privacy* (protezione dal rischio di identificazione o ricostruzione dei dati sorgente) e *fairness* (assenza di distorsioni discriminatorie o ingiustificate)**. La somiglianza statistica con il dataset sorgente non dimostra che un modello addestrato sui dati sintetici funzioni sul mondo reale; un buon risultato medio può nascondere errori su gruppi, eventi rari o periodi di stress. Le prassi raccolte dal Synthetic Data Expert Group della FCA - utili come riferimento tecnico, pur non costituen-

1 EDPB, Guidelines 02/2026 on Anonymisation, versione 1.0, 7 luglio 2026, in consultazione pubblica fino al 30 ottobre 2026, executive summary e parr. 45-53; cfr. CGUE, 4 settembre 2025, C-413/23 P, EDPS c. SRB.

2 EDPB, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, 17 dicembre 2024, parr. 34, 38 e 43; T. Stadler, B. Oprisanu, C. Troncoso, "Synthetic Data - Anonymisation Groundhog Day", USENIX Security 2022.

3 Regolamento (UE) 2024/1689, come modificato dal Regolamento (UE) 2026/1744, artt. 4-bis, 10 e 113 e Allegato III, punto 5, lett. b) e c). Nel testo originario la base per il trattamento delle categorie particolari era contenuta nell'art. 10, par. 5; la modifica del 2026 la ha trasferita e ampliata nell'art. 4-bis.

4 EIOPA, Opinion on Artificial Intelligence Governance and Risk Management, EIOPA-BoS-25-360, 6 agosto 2025, parr. 3.11, 3.19, 3.21 e 3.23; BCE, Pedro Machado, "Technology is neutral, governance is not: AI adoption in the banking sector", 24 febbraio 2026.

do propriamente una guidance - raccomandano provenienza, gestione delle versioni, test avversariali, pluralità di metriche e monitoraggio nel tempo.[5]

Si pensi a un portafoglio sinistri sintetico usato per addestrare un modello di *triage*, ovvero quel modello che classifichi preliminarmente i sinistri. Se il generatore attenua la frequenza dei grandi sinistri, il modello può apparire accurato e, allo stesso tempo, sottostimare proprio le esposizioni più rilevanti. Servono un **campione di dati reali mai utilizzati dal generatore**, test sulle code, analisi temporali e per sottogruppi, nonché il *parallel running* prima dell'uso produttivo (prevedendo quindi il funzionamento in parallelo rispetto al sistema già in uso).

Va tracciata anche la catena delle generazioni. La ricerca sul "*model collapse*", cioè sul degrado progressivo dei modelli addestrati ricorsivamente su dati generati da altri modelli, mostra che l'addestramento ricorsivo e indiscriminato su dati prodotti da modelli può progressivamente perdere le code della distribuzione. Il risultato riguarda specifici scenari di apprendimento ricorsivo e non ogni dataset sintetico; è comunque un motivo concreto per registrare se e quante volte dati sintetici siano divenuti input di generatori successivi.[6]

5. Il "passaporto" del dataset nella due diligence M&A

La *data room* dovrebbe contenere un "passaporto" del *dataset* per ogni *asset* materiale. Con questa espressione si intende un fascicolo tecnico e giuridico che consenta al compratore di ricostruire l'origine dei dati, il metodo con cui sono stati generati, i diritti che ne permettono l'utilizzo, le verifiche effettuate e le condizioni necessarie per riprodurli e aggiornarli. Non è quindi una semplice scheda descrittiva: serve a verificare che il dataset conservi valore e utilizzabilità dopo l'operazione e a tradurre gli esiti della due diligence in garanzie, condizioni e obblighi contrattuali. Il documento può essere organizzato in sei livelli.

5 FCA Synthetic Data Expert Group, "Generating and using synthetic data for models in financial services: governance considerations", 2025, spec. parr. 1.8, 3.19-3.23, 3.36-3.42 e 3.46-3.56. Il report precisa di non costituire guidance o posizione della FCA.

6 I. Shumailov et al., "AI models collapse when trained on recursively generated data", Nature, vol. 631, 2024, pp. 755-759, doi:10.1038/s41586-024-07566-y.

1. **Origine e diritti:** categorie e periodo dei dati sorgente, modalità e finalità di raccolta, basi giuridiche e informative privacy, licenze, restrizioni d'uso e di text and data mining, fornitore dei dati e consensi al cambio di controllo.

2. **Genealogia tecnica:** generatore, versione, codice, parametri, *seed* (il valore iniziale che orienta la generazione pseudocasuale e contribuisce a renderne ripetibile il risultato), trasformazioni, percentuale reale/sintetica, regole di arricchimento, dipendenze cloud e componenti di terzi. La conservazione del solo seed non è sufficiente: la generazione deve poter essere replicata in un ambiente pulito utilizzando la medesima versione del codice, le stesse configurazioni e le dipendenze necessarie.

3. **Evidenze di qualità:** metriche e soglie per ciascuna specifica finalità d'uso (*intended purpose)*, confronto con dati reali tenuti separati, analisi di code e sottogruppi. Occorre ottenere risultati negativi e interventi correttivi, non solo demo selezionate.

4. **Evidenze privacy e sicurezza:** metodo di anonimizzazione, *threat model* (analisi dei soggetti, delle tecniche e degli scenari di attacco), attacchi simulati, accessi al generatore, *data leakage* (fuoriuscita o ricostruzione indebita dei dati sorgente), incidenti, reclami, richieste degli interessati e interlocuzioni con autorità.

5. **Trasferibilità:** titolarità e licenze su output, generatore, documentazione e know-how; diritti di modifica e sublicenza; portabilità di configurazioni e log; *escrow* o continuità per componenti essenziali. L'investimento nella creazione dei record non garantisce da solo la tutela sui generis della banca dati; assumono quindi peso concreto contratto, segreto commerciale e misure adottate per conservarlo.[7]

6. **Governance:** *data owner* e *model owner*, approvazioni, controlli indipendenti, registro delle modifiche (*change log*), frequenza di rivalutazione, usi vietati e processo di ritiro.

7 Direttiva 96/9/CE, art. 7; CGUE, 9 novembre 2004, C-203/02, British Horseracing Board, parr. 31-36 e 42 (distinzione tra investimento per ottenere contenuti esistenti e risorse impiegate per crearli); Direttiva (UE) 2016/943, art. 2, n. 1, sui segreti commerciali.

vedi l'articolo online

La struttura dell'operazione incide direttamente sul rischio. In uno **share deal** (volto all'acquisto di partecipazioni) cambia il controllo della target, ma la società rimane la stessa: dataset, generatore, codice, personale e contratti restano, in linea di principio, nella titolarità della target. Ciò non elimina la necessità di verificare le clausole di *change of control* ~~(cambio di controllo)~~ contenute nelle licenze, nei contratti cloud e negli accordi di fornitura dei dati, che possono richiedere un consenso, consentire il recesso o modificare le condizioni economiche. Occorre inoltre accertare se le competenze necessarie a mantenere il generatore dipendano da personale chiave (che quindi dovrà essere destinatario di specifica regolamentazione nel contratto) e se l'integrazione della target nel gruppo comporti nuovi accessi ai dati, nuove condivisioni o una diversa finalità d'uso. Anche se l'identità del titolare del trattamento normalmente non cambia per il solo acquisto delle partecipazioni, l'utilizzo dei dati da parte di altre società del gruppo deve avere un autonomo fondamento giuridico ed essere coerente con le finalità originarie.

In un **asset deal** (volto all'acquisto dell'azienda, di un ramo o di singoli beni e rapporti), invece, il compratore deve verificare, per ciascun elemento della catena, se rientri effettivamente nel perimetro dell'acquisizione e se sia trasferibile. Il perimetro dovrebbe comprendere espressamente i dati sintetici e, quando necessari, i dati sorgente, il generatore, il codice, i parametri, la documentazione, i diritti di proprietà intellettuale, le licenze, i contratti, i registri di validazione e le risorse o i servizi indispensabili per continuare a utilizzare e aggiornare il dataset. Devono inoltre essere ottenuti gli eventuali consensi di terzi e verificata la possibilità giuridica di trasferire e riutilizzare i dati. Se il compratore commercializza il sistema con il proprio marchio, lo modifica sostanzialmente o ne cambia la finalità in modo da renderlo ad alto rischio, può assumere il ruolo di provider (fornitore ai sensi dell'AI Act) e i relativi obblighi.

6. Dalla due diligence allo SPA

Le criticità emerse devono produrre conseguenze verificabili nello SPA. Le **dichiarazioni e garanzie** dovrebbero coprire completezza dell'inventario; diritti e liceità dell'uso dei dati sorgente; correttezza del "passaporto"; riproducibilità delle versioni del software e delle configurazioni utilizzate per la generazione (*build*); assenza di incidenti, contestazioni o indagini non dichiarati; rispetto delle metriche espressamente riportate. Una garanzia assoluta di assenza di bias o di anonimato è difficilmente sostenibile; risultano più efficaci garanzie su processo, evidenze, soglie e informazioni fornite al compratore (*disclosure*).

Il periodo tra *signing* e *closing* consente di colmare le carenze che non possono rimanere aperte al trasferimento. I ***covenant* (impegni che regolano la condotta delle parti tra signing e closing)** possono obbligare il venditore a preservare dataset, generatore, documentazione e personale chiave e a non modificare dati sorgente, parametri o processi di generazione senza il consenso del compratore. Le **condizioni sospensive** possono invece subordinare il closing all'ottenimento dei consensi di terzi, alla consegna del codice e della documentazione mancanti, alla segregazione dei dati privi di un titolo adeguato, al superamento di un test indipendente o alla replica del generatore nell'ambiente del compratore. La distinzione ha un effetto pratico: l'inadempimento di un covenant espone il venditore ai rimedi contrattuali, mentre il mancato avveramento di una condizione sospensiva può impedire il closing, salvo rinuncia della parte nel cui interesse è prevista.

La tecnica di allocazione deve dipendere dalla natura del rischio. Una carenza sanabile può richiedere un intervento correttivo (*remediation)* prima del *closing;* una perdita di valore già quantificabile può riflettersi sul prezzo; un rischio specifico ma incerto può essere coperto da un indennizzo e, se opportuno, da un *holdback* (trattenuta di parte del prezzo) o da un *escrow* (deposito della somma presso un terzo). Se l'incertezza riguarda le prestazioni future, la componente variabile del prezzo (*earn-out*) dovrebbe dipendere da metriche concordate, misurate su dati reali tenuti separati e non modificabili unilateralmente dal venditore. Il contratto dovrebbe inoltre indicare chi prepara i dati di verifica, chi esegue il test, come sono risolte le divergenze metodologiche e quali conseguenze derivano dal mancato raggiungimento delle soglie.

Esempio: una target dichiara che il modello di credito è stato addestrato con il 60% di record sintetici per correggere una popolazione sottorappresentata. Prima del closing, il compratore verifica performance complessiva, falsi negativi, disparità tra gruppi, stabilità temporale e leakage. **Soglie, dataset di verifica, conseguenze del mancato superamento e diritto di ripetizione del test** devono essere definiti prima che il risultato sia noto.

7. L'acquisto di soluzioni AI da parte di banche e assicurazioni

Nel procurement, il "passaporto" diventa un allegato contrattuale e un deliverable aggiornabile. La dichiarazione del fornitore secondo cui la soluzione utilizza "solo dati sintetici" non consente alla banca o all'assicurazione di governare il rischio. Per i sistemi rientranti nella sua Opinion del 2025, EIOPA richiede informazioni e evidenze su caratteristiche, capacità, dati di training e test e limitazioni; la protezione dell'IP del fornitore può essere salvaguardata con accessi controllati, esperti indipendenti e report graduati, senza svuotare audit e controllo.

Quando la soluzione è un servizio ICT che supporta una funzione critica o importante, DORA richiede due diligence precontrattuale, diritti di accesso, audit e ispezione, gestione della subfornitura, disponibilità e restituzione dei dati ed *exit strategy*. **La natura sintetica del dataset non elimina dipendenza dal fornitore, lock-in o rischio di concentrazione.**[8]

Anche i ruoli AI Act vanno anticipati. Un *deployer* può diventare provider di un sistema ad alto rischio quando appone il proprio marchio, effettua una modifica sostanziale o ne cambia la finalità così da renderlo ad alto rischio. L'articolo 25, come modificato nel 2026, dettaglia la cooperazione dovuta dal provider iniziale al nuovo provider: documentazione tecnica, limiti e modalità di malfunzionamento conosciute noti e accesso tecnico mirato, anche per test e validazione.[9] Il contratto dovrebbe assicurare questi elementi prima che una personalizzazione produca il passaggio di ruolo.

L'allegato dati dovrebbe quindi fissare almeno: **usi consentiti; origine e percentuale dei dati sintetici; versione del generatore; metriche di privacy, qualità e *fairness; benchmark* di accettazione su dati reali; frequenza dei test; eventi di modifica materiale; diritti di audit; divieto di addestrare modelli generali con dati, *prompt* (istruzioni impartite al modello), *log* (registri delle interazioni) o *feedback* del cliente senza autorizzazione; supporto regolatorio; portabilità ed exit** (se non già disciplinati nel corpo principale del contratto). La sostituzione del dataset sorgente, del generatore o dei parametri privacy e un cambiamento significativo della quota sintetica dovrebbero attivare notifica, rivalutazione e, per gli usi più sensibili, approvazione preventiva.

Un caso operativo è l'acquisto di una soluzione AML alimentata con scenari sintetici. La FCA ha sviluppato nel 2026, a fini di ricerca, un dataset interamente sintetico derivato da dati di retail banking e arricchito con scenari realistici di riciclaggio: un esempio del potenziale della tecnica, non una prova automatica di efficacia.[10]

In gara, tutti i fornitori dovrebbero essere valutati nelle stesse condizioni e su un insieme di prova non conosciuto in anticipo, composto, ove possibile e lecito, da operazioni reali adeguatamente protette e da casi validati da esperti. Prima della prova occorre definire metriche e soglie comuni: capacità di individuare le operazioni sospette, falsi positivi e falsi negativi, efficacia sulle tipologie rare, stabilità dei risultati e tempi di elaborazione. L'insieme di prova non dovrebbe essere utilizzabile dal fornitore per riaddestrare o calibrare il sistema, perché in tal caso il confronto misurerebbe l'adattamento alla prova anziché la capacità effettiva della soluzione. Per rendere comparabili le offerte, dovrebbero inoltre essere uniformi il periodo osservato, le informazioni disponibili al sistema e le regole con cui gli esiti sono classificati come corretti o errati.

Dopo l'aggiudicazione, la *shadow mode* - ovvero la fase di funzionamento parallelo senza effetti automatici sulle decisioni operative - consente al nuovo sistema di analizzare operazioni reali e generare segnalazioni mentre gli operatori continuano a utilizzare il sistema esistente. Gli esiti dei due sistemi possono così essere confrontati, gli errori esaminati e le soglie calibrate prima dell'avvio produttivo. Il contratto dovrebbe stabilire la durata e i criteri di uscita da questa fase, le responsabilità nell'analisi degli scostamenti, gli obblighi di monitoraggio e il diritto della banca di contestare o rifiutare aggiornamenti che alterino metriche, soglie o prestazioni concordate. Il passaggio alla produzione dovrebbe avvenire soltanto dopo il superamento dei criteri di accettazione e l'approvazione delle funzioni interne competenti.

8 Regolamento (UE) 2022/2554 (DORA), in particolare artt. 28 e 30, sugli accordi con fornitori terzi di servizi ICT e sui servizi a supporto di funzioni critiche o importanti.

9 Regolamento (UE) 2024/1689, art. 25, come modificato dall'art. 1 del Regolamento (UE) 2026/1744; cfr. in particolare i parr. 1, 2 e 4 sul passaggio di ruolo e sulla cooperazione nella catena del valore.

10 FCA, "Synthetic Data and Anti-Money Laundering - Project Report", Research Note, 15 aprile 2026. La Research Note precisa che i risultati non rappresentano necessariamente la posizione della FCA.

8. Dati sintetici: il valore è nella prova

Nei dati sintetici **il valore coincide con la capacità di ricostruirne la genealogia e dimostrarne l'idoneità**. Per un acquirente, l'asset comprende il dataset, il generatore, i diritti, la documentazione, le prove e le persone e infrastrutture necessarie a riprodurlo. Per una banca o assicurazione che acquista una soluzione AI, gli stessi elementi devono diventare obblighi del fornitore, criteri di accettazione e controlli lungo il ciclo di vita.

Il "passaporto" del dataset crea un collegamento operativo tra data science, privacy, IP, AI Act, disciplina prudenziale e DORA. Consente di trasformare un'affermazione commerciale - "i dati sono sintetici" - in una sequenza di evidenze verificabili. È questa sequenza, più del volume dei record, a determinare **valore, governabilità regolatoria e continuità operativa**.

vedi l'articolo online

DB non solo diritto bancario

A NEW DIGITAL EXPERIENCE

dirittobancario.it