

Di Enrico Nerli Ballati

I. La definizione di analisi secondaria

«Con analisi secondaria si intende l'analisi dei dati precedentemente raccolti, disponibili e organizzati in modo sistematico rispetto a una unità di analisi individuale o aggregata, provenienti da una o più fonti statistiche, con lo scopo di rispondere a una domanda di ricerca definita indipendentemente dalle finalità per cui i dati sono stati originariamente raccolti» (Biolcati-Rinaldi, Vezzoni, anno, p. 16). La definizione proposta contiene al suo interno una serie di elementi costitutivi che è possibile sviluppare in termini analitici (*ivi*, pp. 16-19).

1. *Si tratta di un'analisi di dati già raccolti e disponibili.* Caratteristiche fondamentali e trasversali a tutte le definizioni di analisi secondaria presenti in letteratura.

2. *Organizzati in modo sistematico.* Le informazioni dovrebbero essere già disponibili in forma organizzata per una potenziale analisi. Il dataset da analizzare dovrebbe già esistere materialmente.

3. *Rispetto a una specifica unità di analisi individuale o aggregata.* L'analisi si può riferire sia a unità di analisi costituite dagli individui, sia a unità di analisi aggregate (es. unità amministrative quali comuni, province, regioni). Nella definizione di analisi secondaria non è fondamentale la definizione dell'unità di analisi, quanto che le informazioni siano organizzate e disponibili sistematicamente in relazione a una determinata unità di analisi.

4. *Provenienti da una o più fonti statistiche.* Il caso più semplice è l'analisi dei dati originati da una singola fonte, ad esempio un'indagine diretta con questionario standardizzato che ha coinvolto come unità di rilevazione degli individui. Nello stesso tempo i dati originati da più indagini dirette o fonti amministrative possono essere combinati. Vedremo in seguito come questa possibilità sia fondamentale nella valorizzazione delle fonti originate da dati amministrativi e per definire lo stesso concetto di qualità del dato in termini di coerenza dei dati.

5. Con lo scopo di rispondere a una domanda di ricerca definita indipendentemente dalle finalità per cui i dati sono stati originariamente raccolti. L'analisi secondaria si configura come una nuova attività di ricerca nella quale viene formulata una nuova domanda di ricerca, indipendentemente dagli obiettivi cognitivi che hanno guidato la progettazione del processo di produzione dei dati originario.

Per il ricercatore impegnato nell'attività scientifico sociale, l'analisi secondaria presenta molteplici vantaggi. In primo luogo, muovere nella ricerca e analisi di fonti secondarie consente un risparmio notevole di costi in termini di tempo, di organizzazione del lavoro, di risorse umane ed economiche che qualsiasi processo di progettazione degli strumenti e di raccolta dei dati quantitativi comporta. Questa opportunità diventa ancor più rilevante qualora il disegno di ricerca si sviluppi in un orizzonte temporale di breve/medio periodo e coinvolga un gruppo di ricerca limitato o addirittura un solo studioso. Esempificativo a tal proposito il caso del lavoro per le analisi empiriche connesse alla redazione delle tesi di laurea o di dottorato (*ivi*, p. 29).

Un secondo vantaggio fondamentale risiede nella capacità dell'analisi secondaria di configurarsi come strumento utile allo studio dei fenomeni sociali situati nel passato, rispetto ai quali non è possibile produrre nuove informazioni attraverso indagini dirette. In questo senso, ricorrere a fonti di carattere secondario, formulando nuove domande alle quali rispondere attraverso dati raccolti per altre finalità, rappresenta l'unica strategia percorribile. Esempi chiari e articolati sono rinvenibili nell'ampio archivio dei dati delle serie storiche¹ dell'Istituto Nazionale di Statistica (Istat, 2011; Istat, 2022). Ad esempio, partendo da un atto amministrativo quale la sottoscrizione dell'atto di matrimonio da parte degli sposi, è stata elaborata attraverso la Rilevazione mensile degli eventi di stato civile la serie storica dell'indicatore *Sposi che non sottoscrissero l'atto di matrimonio perché non sapevano scrivere* relativa all'intervallo temporale 1867-1965. Il dato è stato originato da un atto amministrativo che non ha finalità dirette di ricerca, ma l'informazione sulla capacità di sottoscrizione dell'atto fu raccolta a partire dal 1867 dalla Rilevazione mensile degli eventi di stato civile (avviata nel 1862), assumendo quindi la rilevanza pubblica e scientifica di questa determinata informazione statistica (Istat, 2011). Ai giorni nostri, questi dati costituiscono un indicatore del tema *Istruzione* all'interno della macroarea *Istruzione e lavoro* nell'ambito della classificazione degli indicatori delle Serie storiche Istat (Istat, 2022). L'indicatore consente di osservare l'andamento dell'analfabetismo in Italia, anche ponendolo in relazione con determinate fasi storiche o con l'implementazione di riforme o di politiche dell'istruzione nel passato. Parimenti, la possibilità di effettuare elaborazioni sui dati aggregati consente ai ricercatori di ricostruire misure sulle differenze di genere nell'analfabetismo, comparando l'andamento del dato femminile e del dato maschile, o di rappresentare i divari territoriali nell'istruzione, analizzando le di-

¹ <https://seriestoriche.istat.it/>

stanze relative per l'andamento dell'indicatore tra le varie ripartizioni geografiche o tra le regioni del Paese.

Un terzo vantaggio rilevante nel campo della ricerca sociale risiede nella possibilità di disporre attraverso l'analisi secondaria di definizioni operative e quindi di dati, in grado di rispondere alle caratteristiche di *coerenza* e *confrontabilità* che costituiscono alcune tra le dimensioni del concetto di *qualità* del dato statistico (Eurostat, 2003; Istat, 2016; Eurostat, 2021). Nella ricerca sociologica un problema ricorrente per i disegni di ricerca che prevedano la progettazione degli strumenti e la raccolta dei dati ex novo è la produzione di definizioni operative e di rappresentazioni dei fenomeni oggetto di studio difformi o addirittura pleonastiche, rispetto a quanto già presente e sedimentato nella letteratura scientifica. Se da un lato, nelle indagini campionarie, la formulazione e riformulazione dei quesiti con la conseguente creazione di nuove variabili sui fenomeni oggetto di studio costituisce un esercizio utile ad adattare gli strumenti di rilevazione ai singoli contesti e a stimolare la riflessione metodologica, dall'altro la proliferazione delle definizioni operative mina la coerenza e la confrontabilità dei risultati dello studio. In questo senso oltre ai vantaggi inerenti al risparmio di tempi e risorse, il ricercatore o il gruppo di ricerca che si avvicina ai dati secondari prodotti da enti di ricerca beneficia di una sorta di *vincolo esterno*, soprattutto quando questi dati costituiscano delle statistiche ufficiali rispondenti a precisi standard di qualità.

La *coerenza* indica l'adeguatezza delle statistiche ad essere combinate in modo diverso e per diversi usi. Si parla di riconciliabilità tra statistiche all'interno di una stessa fonte relative a variabili diverse, calcolate su domini diversi, da fonti diverse o da processi con periodicità diverse (Istat, 2016, p. 40). Mentre la *comparabilità* nel tempo e geografica indica quanto le differenze nel tempo e tra aree geografiche siano dovute a variazioni reali del fenomeno oggetto di studio e non a differenze in concetti statistici, strumenti di misurazione e procedure (*ibidem*). A titolo di esempio, il ricercatore interessato allo studio dell'inserimento professionale e alla stratificazione sociale della popolazione giovanile in una determinata area del Paese e in un determinato periodo potrà ricorrere a dati secondari di statistica ufficiale raccolti presso individui e famiglie, che consentono di organizzare le informazioni sul lavoro svolto dai giovani secondo la *Classificazione delle Professioni* dell'Istat CP2011. Questa classificazione riconduce la professione svolta dall'individuo a determinati raggruppamenti professionali organizzati in maniera gerarchica a diversi livelli di analiticità, dal primo livello di massima sintesi – composto da 9 grandi gruppi professionali – al quinto livello e ultimo livello di classificazione, composto da 800 unità professionali (Istat, 2013). I codici relativi alla classificazione CP2011 sono a loro volta riconducibili alla *International Standard Classification of Occupations – Isco08*, ampiamente condivisa a livello internazionale. Seguendo procedure metodologiche pubbliche e replicabili, la disponibilità delle informazioni sui codici

Isco08 sulla professione degli individui consentirà di riprodurre classificazioni presenti in letteratura per soddisfare gli obiettivi cognitivi dello studio, come ad esempio lo schema delle classi sociali dell'*European Socio-Economic Classification*, basato su diverse procedure di ricodifica delle professioni rappresentate dalla classificazione *Isco08* (Rose e Harrison, 2007). Evidenti in questo senso i vantaggi in termini di potenziale comparabilità dei risultati dell'analisi secondaria effettuata, in termini di connessioni con altri studi presenti in letteratura e di possibilità di operare un confronto spaziale e temporale dei dati. Possibilità che qualificerebbero positivamente lo studio, anche qualora l'analisi secondaria funga da semplice punto di partenza per eventuali approfondimenti qualitativi o quantitativi sul fenomeno in uno specifico contesto. Si tratta di risultati che si raggiungerebbero con oneri e rischi metodologici notevolmente maggiori, tentando una rilevazione diretta sulla popolazione. Nello stesso tempo la possibilità di aggregare i dati secondari per unità di analisi territoriali quali comuni, province o regioni, oppure la possibilità di derivare variabili secondo procedure replicabili, consentirà di combinare i dati con altre fonti secondarie provenienti da statistiche ufficiali, costruendo una base dati più ampia per lo studio del fenomeno di interesse. Ad esempio, combinando informazioni sulla domanda e offerta di lavoro, informazioni sul funzionamento del sistema di istruzione e di formazione o sui livelli di competenze della popolazione scolastica prima dell'entrata nel mercato del lavoro nel territorio e per il periodo di interesse per l'analisi.

2. Indagini dirette, fonti amministrative e sistemi informativi statistici

L'ampia mole di dati secondari disponibile può essere originata da differenti fonti e organizzata in molteplici modalità. Nell'ambito del linguaggio della statistica ufficiale italiana viene impiegata una classificazione che ne riduce la complessità e riconduce i prodotti disponibili ad alcune categorie (Sistan, Istat, 2018). Tra queste categorie, una prima definizione utile è quella di Statistiche da indagine. Si tratta di «processi di produzione di informazioni statistiche attraverso la rilevazione diretta da unità rispondenti (soggetti pubblici o privati, individuali o collettivi). Le informazioni statistiche sono acquisite presso (tutte o parte) le unità di un collettivo, secondo un disegno di indagine di tipo statistico» (*ivi*, p. 144). Rientrano in questa categoria le indagini campionarie svolte con strumenti di rilevazione standardizzati, che vedono come unità di rilevazione l'individuo, che costituiscono uno degli strumenti elettivi della ricerca sociale.

Un esempio di dati secondari originati da indagine diretta è costituito dai dati dell'indagine internazionale *European Social Survey ESS*, condotta per

l'Italia dall'Inapp (Istituto Nazionale per l'Analisi delle Politiche Pubbliche), che rileva gli atteggiamenti, le credenze e i comportamenti dei cittadini di più di trenta nazioni. La ESS mira a: registrare stabilità e cambiamenti delle situazioni sociali e degli atteggiamenti in Europa, nonché interpretare il cambiamento del tessuto sociale, politico e morale degli europei; introdurre indicatori sensibili per rilevare le dinamiche nazionali, basati sulle percezioni e le valutazioni dei cittadini relative ad aspetti chiave del loro Paese; migliorare la visibilità e la diffusione dei dati sul cambiamento sociale fra gli studiosi, i policy makers e, più in generale, nell'opinione pubblica (Sistan, Istat, 2022a, p. 7). Un altro, tra i tanti esempi disponibili, è costituito dai dati dell'indagine internazionale *Programme for International Student Assessment PISA*, promossa dall'Ocse e condotta in Italia dall'Invalsi, che valuta con periodicità triennale il livello di istruzione degli adolescenti dei principali Paesi industrializzati e in che misura gli allievi prossimi alla conclusione dell'obbligo scolastico abbiano acquisito alcune delle conoscenze e delle abilità essenziali del loro corso di studi (Sistan, Istat, 2022a, p. 46). Informazioni corredate da dati articolati sull'origine sociale e le condizioni socio-economiche e culturali degli studenti. Entrambe le indagini sono svolte nell'ambito del Programma statistico nazionale (*cf.* box 17.2). Rispetto a questo tipo di indagini, il ricercatore potrà acquisire sia dei macrodati aggregati, in forma di indicatori su specifiche unità geografiche o classi di soggetti, sia dei microdati, vale a dire collezioni di dati elementari riferiti alle unità di analisi dello studio sulle quali effettuare analisi personalizzate, che possono condurre anche a nuovi risultati originali.

Negli anni recenti la produzione di dati a partire da archivi amministrativi ha assunto valenze sempre più crescenti in Italia. Questo in ragione dell'opportunità di valorizzare la grande mole di informazioni strutturate e riutilizzabili connesse allo sviluppo dell'informatizzazione nella Pubblica amministrazione, in ragione della necessità di ridurre i costi delle rilevazioni dirette e di integrarne i risultati con i dati amministrativi e della possibilità di ridurre gli oneri a carico dei rispondenti (cosiddetto fastidio statistico). Nell'ambito della classificazione proposta dal Sistan, possiamo definire le Statistiche da fonti amministrative organizzate (Sda) nei termini di «processi di produzione di informazioni statistiche attraverso un processo di trasformazione condotto su fonti amministrative organizzate, pubbliche o private (registri, archivi, basi di dati). Le informazioni statistiche derivano, in questo caso, da un processo volto a garantire sia diversi aspetti della qualità dell'informazione statistica sia gli aspetti relativi alla tutela della riservatezza di informazioni raccolte per fini amministrativi» (Sistan, Istat, 2018, p. 144). Un esempio di statistica da fonti amministrative è rappresentato dai dati dello studio "Cittadini non comunitari regolarmente presenti in Italia" dell'Istat, che, attraverso l'acquisizione dei dati dell'Archivio dei permessi di soggiorno del Ministero dell'Interno, permette di analizzare la presenza straniera regolare, mettendo in luce le principali differenze di condizioni e i comportamenti degli stranieri in base ad alcune caratteristiche socioanagrafiche (età, sesso, paese di nascita, cittadinanza, stato civile, motivo del rilascio del permesso), consentendo inoltre di approfondire il motivo della presenza in Italia e la tipologia di soggiorno (stagionale, a termine, di

lungo periodo). I dati sono a loro volta integrati assieme ad altre fonti statistiche di origine amministrativa nel più ampio progetto “*La presenza straniera in Italia: le condizioni di vita attraverso l'integrazione di archivi amministrativi*” dell'Istat, finalizzato a fornire un quadro multidimensionale sui percorsi migratori e di inserimento sociale della popolazione straniera nel nostro Paese (DPR 9 marzo 2022).

Una terza opportunità per l'accesso diretto ai dati secondari, estremamente interessante per il ricercatore, alla luce delle modalità di utilizzo semplificate dei dati, è rappresentata da un Sistema informativo statistico (Sis). Si tratta di un «insieme di informazioni statistiche (dati aggregati e metadati), diffuse su supporti digitali e rese disponibili secondo modalità definite dall'utente, derivanti dall'integrazione concettuale e funzionale di una pluralità di fonti informative (dati elementari e/o aggregati)» (Sistan, Istat, 2022b, p. 15). Un Sis dovrebbe tentare di raggiungere un'elevata esaustività tematica rispetto alle singole fonti potenzialmente utilizzabili e un alto livello di accessibilità ai dati (tale da permettere un'interrogazione personalizzata da parte degli utenti) che si traduce necessariamente nell'impostazione di un Sis su supporti digitali (prevalentemente in ambiente web ma, anche, su supporti digitali che assicurino questo elevato livello di accessibilità) (*ibidem*).

Esempi di sistemi informativi sono il Sistema informativo sui giovani² o Sistema informativo sugli anziani³ realizzati dall'Istat. I dati consultabili sono organizzati per temi, argomenti e indicatori. Si tratta di dati originati da molteplici fonti. Il ricercatore può interrogare il sistema informativo, sia per tema sia per parola chiave, realizzando delle tavole multidimensionali personalizzate su dati aggregati, esportabili in diversi formati informatici, con le quali è possibile descrivere i fenomeni oggetto di interesse e pervenire direttamente a rappresentazioni delle relazioni tra variabili. Ad esempio, realizzando delle tabelle del fenomeno di interesse per area geografica, anno o caratteristiche socio-anagrafiche della popolazione.

Box 17.1 Il Sistan

Il Sistema statistico nazionale (Sistan) è la rete di soggetti pubblici e privati che fornisce al Paese e agli organismi internazionali l'informazione statistica ufficiale (D.Lgs. 322/1989).

Il Sistan comprende: l'Istituto nazionale di statistica (Istat); l'Istituto nazionale per l'analisi delle politiche pubbliche (Inapp, ente d'informazione statistica); gli uffici di statistica delle amministrazioni dello Stato e di altri enti pubblici, degli Uffici territoriali del Governo, delle Regioni e Province autonome, delle Province, delle Camere di commercio (Cciaa), dei Comuni, singoli o associati, e gli uffici di statistica di altre istituzioni pubbliche e private che svolgono funzioni di interesse pubblico. Il progressivo inserimento dei soggetti facenti parte del Sistan è disposto attraverso Decreto del Presidente del Consiglio dei ministri.

² <http://dati-giovani.istat.it/>.

³ <https://dati-anziani.istat.it/>.

Box 17.2 Il Programma Statistico Nazionale (Psn)

«Il Programma statistico nazionale (Psn) è l'atto normativo attraverso cui viene pianificata l'offerta di informazione statistica ufficiale che, nel quadro di coordinamento assicurato dall'Istat, da un lato impegna il Sistema statistico nazionale all'esecuzione dei lavori e progetti statistici di rilievo nazionale ed europeo in esso accolti e, dall'altro, sancisce gli obblighi di collaborazione per i diversi soggetti coinvolti (cittadini, imprese, istituzioni) affinché tali lavori vengano realizzati con gli elevati standard di qualità stabiliti dai codici statistici europeo e nazionale» (Sistan, Istat, 2018).

Al Programma statistico nazionale partecipano l'Istat e altri soggetti titolari di lavori statistici, appartenenti a ministeri, enti e amministrazioni pubbliche centrali, enti locali e altri soggetti. A titolo esemplificativo, gli 811 lavori programmati nel Psn 2020-2022 si articolano nelle aree tematiche: Popolazione e famiglia; condizioni di vita e partecipazione sociale; Salute, sanità e assistenza sociale; Istruzione e formazione; Lavoro e sistema dei trasferimenti monetari previdenziali e assistenziali; Giustizia e sicurezza; Industria costruzioni e servizi: statistiche strutturali e trasversali; Indicatori congiunturali del commercio e altri servizi non finanziari; Pubblica amministrazione e istituzioni private; Ambiente e territorio; Trasporti e mobilità; Turismo e cultura; Agricoltura, foreste e pesca; Statistiche sui prezzi; Indicatori e metodologie per la valutazione delle *policy* (Sistan, Istat, 2022).

Per il ricercatore sociale il Psn rappresenta quindi il luogo naturale in cui esplorare l'ampio panorama di statistiche di interesse pubblico alla ricerca di dati secondari.

3. Accessibilità e chiarezza dei dati secondari

Utilizzando il lessico della statistica ufficiale italiana, nel momento in cui l'utente acquisisce dei dati secondari ci troviamo di fronte a una forma di *diffusione* o di *comunicazione* del dato. I dati sono diffusi quando se ne dà conoscenza a soggetti indeterminati, in qualunque forma, anche mediante la loro messa a disposizione o consultazione. Un esempio di diffusione è rappresentato dalla possibilità di acquisire liberamente tavole statistiche, interrogare e scaricare dati da sistemi informativi statistici, oppure scaricare database in formati elettronici (es. Excel, SAS, SPSS, ecc.), immediatamente fruibili per le analisi. Rispetto alla diffusione dei dati particolare rilievo nel campo degli enti pubblici assume il concetto di dati di tipo aperto (cfr. box 17.3).

Per comunicazione si intende invece il rilascio dei dati a uno o più soggetti determinati (Sistan, 2022, pp. 9-10). Esempi di comunicazione dei dati sono l'acquisizione da parte di un ricercatore di un determinato file di micro-dati attraverso la registrazione presso una piattaforma gestita da un ente/centro di ricerca, oppure l'acquisizione da parte del ricercatore/gruppo di ricerca di file di micro-dati o di elaborazioni specifiche, da effettuare attraverso una richiesta motivata rivolta ad enti o centri di ricerca. Nel caso di Enti e uffici del Sistema

statistico nazionale, le condizioni di base per l'accesso per fini scientifici ai dati elementari dei ricercatori sono definite attraverso apposite Linee guida (Comstat, 2018). In funzione delle finalità del loro impiego e dei bisogni informativi ai quali rispondono, questi file di micro-dati sono classificati in diversi tipi e, qualora si tratti di informazioni su individui, sono sottoposti a misure di tutela della riservatezza finalizzate a non rendere identificabili direttamente o indirettamente i singoli soggetti. In qualsiasi caso, nell'approcciarsi a dati che possano consentire potenzialmente, anche indirettamente, l'identificazione degli individui, il ricercatore e il gruppo di ricerca dovranno aver acquisito familiarità con i concetti alla base della disciplina di protezione dei dati personali (cfr. box 17.4).

Le condizioni di accesso ai dati costituiscono esse stesse una dimensione della qualità del dato statistico, definita come *accessibilità* (Eurostat, 2021). L'accessibilità delle statistiche, infatti, indica la facilità con cui gli utenti possono ottenere i dati. Questa è determinata dalle condizioni attraverso cui gli utenti acquisiscono i dati: dove recarsi, come richiederli, in che tempi di consegna, politica dei prezzi, disponibilità di micro o macrodati, formati disponibili per i dati (Istat, 2016, p. 40).

Accanto alla dimensione dell'accessibilità gioca un ruolo centrale la dimensione della *chiarezza* dei dati da acquisire. La chiarezza delle statistiche è rappresentata dalla facilità con cui gli utenti vengono messi in grado di comprendere i dati. Rilevante per disporre di dati chiari è che questi siano accompagnati da *metadati* appropriati, che siano fornite informazioni sull'accuratezza dei dati (incluse eventuali informazioni sulle limitazioni d'uso) e fino a che punto venga fornita assistenza aggiuntiva dal produttore del dato all'utente che si appropria al suo utilizzo (*ibidem*).

La presenza di metadati assume valenza basilare. Nell'utilizzo dei dati, il ricercatore dovrebbe poter disporre di un tracciato/libro dei codici, con l'indicazione analitica dell'acronimo (o dell'etichetta), della descrizione, del tipo e del formato delle variabili, così come di indicazione e descrizione delle relative modalità delle variabili. Dovrebbe poter disporre, inoltre, di documenti in grado di descrivere la creazione di eventuali nuove variabili ottenute per derivazione dalle variabili originariamente rilevate e disponibili tra i dati rilasciati dal produttore delle statistiche (es. indici sintetici, ricodifiche delle variabili, particolari classificazioni derivate da più variabili). A ciò si dovrebbe poi aggiungere un'eventuale nota metodologica sul processo di acquisizione dei dati dalle fonti amministrative oppure, qualora si tratti di dati originati da un'indagine diretta sulla popolazione, una nota metodologica sulle strategie di campionamento, sulle tecniche di rilevazione adottate, sull'accuratezza delle stime prodotte, sulla periodicità delle stime, sul riferimento territoriale dei dati e sulle unità di analisi disponibili. Il ricercatore, infine, dovrebbe poter disporre di una copia del questionario impiegato, corredato dalle relative istruzioni messe a disposizione dell'intervistatore oppure dal rispondente in caso di auto-somministrazione. Si tratta di requisiti estremamente rilevanti che, qualora siano raggiunti, pongono nelle condizioni di consolidare l'ispezionabilità della base empirica impiegata nell'analisi secondaria.

In termini più generali, prima di procedere all'analisi e anche al momento

della presentazione dei risultati, il ricercatore dovrebbe considerare con attenzione vantaggi e limiti della fonte/fonti impiegata/e, interrogarsi sugli standard di qualità e verificare se e come i dati secondari da utilizzare siano stati già diffusi e siano legati alla produzione di letteratura scientifica.

In questo senso, l'attività di identificazione, valutazione e acquisizione delle fonti di dati secondari rappresenta un'operazione di ricerca fondamentale e nobile di per sé, che consente di acquisire, oltre alle informazioni statistiche, anche informazioni metodologiche sul processo di produzione del dato, sulle definizioni operative e sui risultati di precedenti analisi, in grado di stimolare e sviluppare la capacità di lettura dei fenomeni oggetto di studio.

Box 17.3 Gli Open Data

quando si parla di Open Data, dati di tipo aperto, si fa riferimento alla pratica di mettere a disposizione dati che possono essere liberamente utilizzati, riutilizzati e redistribuiti, con l'unico vincolo al massimo di attribuzione e condivisione nello stesso modo (Sistan, 2022, p. 8).

Nell'ordinamento italiano sono considerati dati di tipo aperto, i dati che presentano le seguenti caratteristiche:

- 1) sono disponibili secondo i termini di una licenza o di una previsione normativa che ne permetta l'utilizzo da parte di chiunque, anche per finalità commerciali, in formato disaggregato (requisito giuridico);
- 2) sono accessibili attraverso le tecnologie dell'informazione e della comunicazione, ivi comprese le reti telematiche pubbliche e private, in formati aperti, sono adatti all'utilizzo automatico da parte di programmi per elaboratori e sono provvisti dei relativi metadati (requisito tecnico);
- 3) sono resi disponibili gratuitamente attraverso le tecnologie dell'informazione e della comunicazione, ivi comprese le reti telematiche pubbliche e private, oppure sono resi disponibili ai costi marginali sostenuti per la loro riproduzione e divulgazione (requisito economico) (Codice Amministrazione Digitale - CAD, art. 1, comma 1, lettera l-ter).

Box 17.4 I dati personali

Dato personale

«Qualunque informazione riguardante una persona fisica identificata o identificabile («interessato») [...] Si considera identificabile la persona fisica che può essere identificata, direttamente o indirettamente, con particolare riferimento a un identificativo come il nome, un numero di identificazione, i dati relativi all'ubicazione, un identificativo online o a uno o più elementi caratteristici della sua identità fisica, fisiologica, genetica, psichica, economica, culturale o sociale» (Sistan, 2022, p. 5).

Trattamento dei dati personali

«Qualsiasi operazione o insieme di operazioni, compiute con o senza l'ausilio di processi automatizzati e applicate a dati personali o insiemi di dati personali, come la raccolta, la registrazione, l'organizzazione, la strutturazione, la conservazione, l'adattamento o la modifica, l'estrazione, la consultazione, l'uso, la comunicazione mediante trasmissione, diffusione o qualsiasi altra forma di messa a disposizione, il raffronto o

l'interconnessione, la limitazione, la cancellazione o la distruzione» (Sistan, 2022, p. 3).

Particolari categorie di dati (ex “dati sensibili”)

«Sono i dati che rivelano l'origine razziale o etnica, le convinzioni religiose, filosofiche, le opinioni politiche, l'appartenenza sindacale, nonché i dati relativi alla salute o alla vita sessuale. Il Regolamento (UE) 2016/679 (articolo 9) ha incluso nella nozione anche i dati genetici, i dati biometrici e quelli relativi all'orientamento sessuale» (Sistan, 2022, p. 7).

Anonimizzazione

«Trattamento dei dati personali che rende impossibile, in modo dimostrabile, riferire le informazioni a una persona fisica identificata o identificabile e, quindi, impedisce l'identificazione dell'interessato (considerando l'art. 26 del Reg. (UE) 2016/679). La sola cancellazione dei dati anagrafici (nome e cognome) o del codice fiscale dell'interessato non determina, di norma, l'anonimizzazione dei dati. A seguito di tale trattamento, i dati non sono più personali» (Sistan, 2022, p. 8).

Pseudonimizzazione

«Trattamento dei dati personali in modo tale che essi non siano più attribuibili a un interessato specifico senza l'utilizzo di informazioni aggiuntive. Tali informazioni vanno conservate separatamente e sono soggette a misure tecniche e organizzative intese a garantire che tali dati personali non siano attribuiti a una persona fisica identificata o identificabile” (art. 4, par. 1, punto 5), del Regolamento (UE) 2016/679). A seguito di tale trattamento i dati sono ancora dati personali» (Sistan, 2022, p. 8).