# REPOSITORIES DEL FUTURO

*Dorit Raines*

Che cosa, come, e dove dovrebbe conservare la sua produzione di dati e di manufatti digitalmente prodotti l'umanità? Nel passato, vista l'esigua densità dei dati prodotta nei secoli – tralasciando momentaneamente l'eliminazione o scomparsa naturale, voluta o accidentata dei dati e soprattutto il loro veicolo, i manufatti – la conservazione era considerata un bene supremo. Con l'accumularsi di manufatti di ogni genere, soprattutto nella sfera amministrativa, si è progressivamente istituita una cultura dello scarto, e quindi di una selezione basata su utilità e vincoli logici, che tendeva ad invadere oltre alla sfera archivistica anche quella biblioteconomica. A questo punto ci si è interrogati anche sull'identità di chi dovrebbe decidere come e cosa conservare. Nel mondo analogico lo hanno fatto le istituzioni culturali deputate a questo scopo: gallerie, biblioteche, archivi, musei (GLAM). Con l'irruzione del mondo digitale – *in primis* la rete – nella nostra vita, la questione sembra più pressante e complessa. Non si tratta più di una questione semantica: archivi, raccolte, collezioni o biblioteche digitali, né tantomeno di politica di recupero (*retrieval*) a partire di metadati. Oggi siamo di fronte a problemi di spazio di archiviazione, di provenienza come fonte di autenticità, di dati collegati (Linked data), di rapida obsolescenza di formati e di una necessità di aggiornamento dei dati.

Se oggi gli archivi digitali sono per lo più piattaforme di deposito di oggetti digitali e loro metadati (come Europeana o archive.org), la questione dell'interoperabilità ovvero lo scambio di informazioni e servizi con altri sistemi o prodotti per potenziarne l'efficienza e ottimizzare le risorse sembra più urgente che mai alla luce dell'esigenza di archiviare un maggior numero di dati nativi digitali, in parte frutto di progetti di ricerca. Oggi si parla di conservazione digitale "classica" quando trattiamo di oggetti digitali e dei loro metadati (possibilmente facendo parte di Linked Open Data - LOD) e di conservazione "dinami-

ca" quando impieghiamo "repositories", dunque sistemi informativi di gestione di datasets; questi ultimi sono stati sinora usati alla stregua dei classici archivi la cui principale funzione era quella di destinatari passivi e silenti delle versioni finali dei risultati di ricerca pubblicati dai loro utenti senza usare la potenzialità della loro peculiare architettura nella gestione ottimale di grandi volumi di dati[1].

Negli ultimi anni è cresciuta la consapevolezza che i dati di ricerca, soprattutto i datasets prodotti durante lo studio, non siano solo una "commodity", e possano perciò avere il loro prezzo, ma che, se vengono messi a disposizione di tutti, specialmente quelli creati mediante un finanziamento pubblico, potrebbero contribuire all'innovazione tecnologica e alla circolazione delle idee. Partendo dal presupposto che ogni entità giuridica sia proprietaria dei propri dati, la comunità internazionale si è interrogata sul principio di "open access", sulla natura e architettura delle infrastrutture che dovrebbero accogliere i dataset di ricerca, sui protocolli e sugli standards e infine sulla funzionalità dei repositories.

In primis, esiste il problema della proprietà dei dati di ricerca e delle infrastrutture che li immagazzinano. Profondi cambiamenti hanno influenzato l'editoria accademica, ma il processo stesso è rimasto notevolmente stabile. Questo comprende quattro funzioni chiave che hanno accompagnato l'editoria scientifica dal Seicento: registrazione (attribuzione), certificazione (peer review), diffusione (distribuzione, accesso), conservazione (memoria accademica e permanente archiviazione). La valutazione è un'altra funzione che è stata associata alla pubblicazione accademica negli ultimi decenni, in particolare attraverso il Journal Impact Factor, ma il suo ruolo è sempre più contestato. Le tecnologie digitali non interrompono le funzioni editoriali, ma consentono la loro distribuzione tra i diversi attori, e non solo gli editori (nel senso tradizionale del termine)[2].

Ma quello dell'editoria accademica non è l'unico tema quando si parla di repositories. Nel mondo, repositories come Genbank, Worldwide Protein Data Bank (wwPDB), e UniProt o l'europeo ELIXIR nelle Scienze della vita; Space Physics Data Facility (SPDF) di NASA e Set of Identifications, Measurements and Bibliography for Astronomical Data (SIMBAD) per le Scienze spaziali, adottano una politica di data curation qualitativa e sofisticata attraverso l'uso di tools

[1] URL : < https://www.coar-repositories.org/activities/advocacy-leadership/working-group-next-generation-repositories/ >

[2] URL: < https://www.openaire.eu/future-of-scholarly-publishing-and-scholarly-communication >

per accedere a contenuti ricchi e dinamici. Tuttavia, molti datasets importanti riconducibili a scienze di benchmark a basso rendimento non si adattano ai modelli di dati dei repositories nati per scopi speciali. Dal momento che anche questi datasets sono ritenuti importanti, la Commissione europea ha deciso nel 2015 di promuovere la European Open Science Cloud (EOSC), una piattaforma open access cloud per poter consultare tutti i dati prodotti dagli scienziati europei[3]. L'anno successivo sono state proposte delle linee guida di Data Management chiamati i principi FAIR: Findability (i dati e i materiali supplementari hanno metadati sufficientemente ricchi e univoci e un link permanente), Accessibility (metadati e dati sono comprensibili agli umani e alle macchine. I dati sono depositati in un repository attendibile), Interoperability (i metadati usano un formato formale, accessibile, una lingua condivisa e ampiamente applicabile per la rappresentazione della conoscenza) e Reusability (dati e collezioni hanno una licenza d'uso e forniscono accurate informazioni sulla provenienza)[4], per poi essere adottati dal summit del G20 nel settembre del 2016[5]. Secondo la visione della Commissione Ue, l'interoperabilità è la funzione del progetto OpenAIRE che ha il compito di collegare tutte le infrastrutture europee dei repositories per poter attuare la politica di "open access" dei dati di ricerca e delle pubblicazioni[6]. Si tratta di una rete di 34 nodi nazionali che interloquiscono con gli "stakeholders" nazionali, promuovono una serie di seminari, localizzano fondi di finanziamento, aiutano le istituzioni nell'adottare un RDM (Research Data Management) policy e agevolano la disseminazione dei materiali che riguardano le idee, i dibattiti e i suggerimenti[7].

L'implementazione dei principi FAIR richiede accordi globali per garantire la più ampia interoperabilità e riusabilità dei dati, oltre a quelli disciplinari e per confini geografici. E poiché esistono simili iniziative nel mondo – si pensi a NIH Data Commons, the Australian Research Data Commons o alla proposta di istituire l'African Open Science Platform – l'idea di rendere i repositories interoperabili a li-

[3] URL: < https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud >.

[4] M.D. Wilkinson et al., The FAIR Guiding Principles for scientific data management and stewardship, «Scientific Data» 3 (2016), URL: < https://www.nature.com/articles/sdata201618 >.

[5] URL: < http://europa.eu/rapid/press-release_STATEMENT-16-2967_en.htm >.

[6] URL: < https://www.openaire.eu/about >.

[7] URL: < https://www.cineca.it/it/progetti/open-access-infrastructure-research-europe-2020 >.

vello globale richiede una serie di sinergie nello sviluppo degli standards e di Data Management Plan (DMP) che COAR (Confederation of Open Access Repositories)[8] si è impegnata ad implementare. Nella visione della COAR i repositories del futuro mireranno a rendere la risorsa – piuttosto che il repository stesso –, l'obiettivo di servizi e infrastrutture. Secondo questa visione, anziché basarsi su metadati descrittivi imprecisi per identificare entità e relazioni tra loro, sarebbe meglio concentrarsi sull'idea intrinseca nell'architettura Web, in cui le entità (note come "risorse") sono accessibili e identificate in modo inequivocabile dagli URL. In questa architettura, sono imprescindibili i riferimenti che vengono copiati tra i sistemi, piuttosto che (come avviene attualmente) i record dei metadati. Inoltre, l'aspettativa degli sviluppatori di repository è quella di automatizzare più possibile l'estrazione di metadati dalle risorse effettive per semplificare il processo di deposito[9].

Infine, esiste la questione delle infrastrutture stesse e dello spazio per immagazzinare o archiviare i dati. A partire del 2005, l'umanità ha superato la soglia analogico/digitale: sono più le informazioni che vengono registrate e conservate utilizzando le tecnologie digitali che quelle analogiche (ad esempio audio nastri, libri, film). Inoltre, il tasso di crescita della generazione digitale si raddoppia ogni qualche anno poiché ora produciamo in un singolo arco di tempo un volume di dati superiore a quello di tutti i periodi precedenti messi assieme. Secondo le stime[10], a partire del 2010 esiste un divario tra la nostra capacità di archiviazione e la generazione dei dati. Stime prudenti suggeriscono che, entro l'anno 2025, le tradizionali tecnologie di memorizzazione dei dati saranno in grado di immagazzinare meno della metà dei dati digitali generati. Aumentare la produzione della tradizionale archiviazione dei dati digitali basata sul silicio non è ritenuta una soluzione sostenibile. La Semiconductor Research Corporation ha previsto che

[8] URL: < https://www.coar-repositories.org/ >.

[9] Le tecnologie e gli standards scelti per semplificare l'interoperabilità sono, tra gli altri: COUNTER, Creative Commons Licenses, ETag, http Signatures, IIIF, Framework, Linked Data Notifications, ORCID, OpenID Connect, ResourcesSync, SUSHI, SWORD, Signposting, Sitemaps, Social Network Identities, Web Annotation Model and Protocol WebID, WebSub e Webmention. URL: < https://www.coar-repositories.org/files/NGR-Final-Formatted-Report-cc.pdf >.

[10] R. Fontana, G. Decad, *Storage Media Overview: Historic Perspectives. Presentation*, May 4, 2016: URL: < http://storageconference.us/2016/Slides/BobFontana.pdf >.

la richiesta di archiviazione di dati digitali supererà di 2-3 ordini di grandezza la fornitura mondiale del silicio entro il 2040[11].

In questo scenario, il DNA rappresenta una tecnologia alternativa per la memorizzazione di informazioni digitali grazie alla capacità di sintetizzazione e alle sue interessanti proprietà.

I computer memorizzano, comunicano e operano su dati binari. Questi bit di informazione sono associati a strutture fisiche e segnali, come lo stato elettronico dei transistor o l'orientamento del campo dei materiali magnetici. La natura archivia anche informazioni digitali, come il codice genetico nelle nostre cellule, sotto forma di polimeri molecolari. Nel DNA, questi polimeri sono costruiti da un insieme di quattro piccole molecole note come nucleotidi (o base). Invece di solo due valori disponibili con dati binari (uno o zero), ogni posizione di nucleotidi nel DNA può assumere uno dei quattro valori (A, C, G o T, che rappresenta il nome chimico della base), quindi ogni base è essenzialmente l'equivalente di informazioni di due bit.

Ogni cellula umana contiene un genoma costituito da circa 6 miliardi di paia di basi di doppio DNA elicoidale, organizzato in 23 gruppi di cromosomi (3 miliardi di coppie di basi di DNA corrisponde ai cromosomi in ciascuna metà del set). Il DNA in questi cromosomi codifica circa 1,6 gigabyte di informazioni in totale, per cellula. Quindi, se si assommano tutte le cellule del nostro corpo, arriveremo a circa 100 zettabytes di informazioni contenute nel nostro DNA.

Tuttavia, il DNA non possiede solo una capacità di archiviazione di una quantità elevatissima di dati: la densità di informazioni che può accogliere è di gran lunga superiore a ogni altra tecnologia, così come il suo modo di archiviare dati su tre dimensioni – quindi la sua capacità volumetrica – il che rende questa tecnologia più efficace poiché risparmia spazio fisico di archiviazione di dati. Un altro motivo per caldeggiare l'uso del DNA come tecnologia di conservazione dei dati è la sua stabilità: è capace a preservare dati per secoli e, in caso di degrado, con i sistemi di ridondanza e di correzione di errori siamo in grado di superare il problema. Infine, copiare il DNA usando procedure già note nel campo della biologia molecolare risulta veloce e a basso costo[12].

[11] V. Zhirnov, R.M. Zadegan, G.S. Sandhu, G.M. Church, W.L. Hughes, *Nucleic acid memory*, «Nature Materials» 15 (2016), 4, pp. 366-370.

[12] Bornholt et al., A DNA-based archival storage system, «IEEE Computer Society» (2017), 3, pp. 98-104: < https://homes.cs.washington.edu/~bornholt/papers/dnastorage-toppicks17.pdf >.

La prima idea di usare il DNA come "piattaforma" di archiviazione dei dati risale al 1995[13], ma i primi esperimenti furono effettuati nel 1999[14]. Dopo qualche anno di stallo dovuto ai problemi di sintetizzazione del DNA e di impiego delle tecniche di sequenziamento, recentemente, e dopo qualche anno di collaborazione tra Microsoft, Università di Washington e Twist Bioscience, i ricercatori sono riusciti a codificare 200 MB di informazione in DNA e recuperare questi dati con 100% di accuratezza[15]. Oggi, le sfide di fronte ai ricercatori sono di tre tipologie: l'aumento della densità di archiviazione, la correzione di errori nella sintesi e nel sequenziamento del DNA e le tecniche di compensazione per il possibile degrado del DNA.

Il DNA potrebbe divenire allora il nostro cloud del futuro? Sebbene il DNA abbia un enorme potenziale come dispositivo di memorizzazione dei dati del futuro, è necessario risolvere molteplici problemi come costi esorbitanti, meccanismi di scrittura e lettura lenti e vulnerabilità (per via di mutazioni o errori)[16]. Con la rapidità degli sviluppi tecnologici è molto probabile che tra qualche anno sapremo se il futuro dell'archiviazione dei dati, inclusi anche i dati della ricerca, sarà affidato al DNA.

[13] E.B. Baum, *Building an associative memory vastly larger than the brain*, «Science» 268, 5210 (1995), 4, pp.583-585.

[14] C.T. Clelland, V. Risca, C. Bancroft, *Hiding messages in DNA microdots*, «Nature» 399, 6736 (1999), 6, pp. 533-534.

[15] DNA-Based Digital Storage, *Twist Bioscience White Paper*, pp. 1-5, URL: < https://pdfs.semanticscholar.org/fbd5/af58f65f69a0661e26a496d1a440979ab25b.pdf >.

[16] D. Panda et al., DNA as a digital information storage device: hope or hype?, «3 Biotech» 8, 5 (2018), pp. 1-9, URL: < https://www.researchgate.net/publication/324945710_DNA_as_a_digital_information_storage_device_hope_or_hype >.