DOI: 10.26314/GARR-Conf22-proceedings 21

Progetto Arkive: un'infrastruttura per l'archiviazione a lungo termine dei dati della ricerca dell'Università degli Studi di Milano

Federica Zanardini

Università degli Studi di Milano – Direzione ICT

Abstract. L'Università degli Studi di Milano, una tra le maggiori Università italiane, produce un vasto numero di contenuti digitali: dati della ricerca, dati e oggetti digitali prodotti a sup-porto delle attività didattiche, digitalizzazioni di beni culturali, archivistici e librari e dati di carattere amministrativo. Un censimento condotto all'inizio del 2021 ha permesso di conoscere l'entità e il vo-lume delle esigenze di storage per l'archiviazione (dell'ordine delle decine di PB con in-crementi annuali stimati di 20PB), nonché di constatare l'estrema frammentazione e di-somogeneità delle soluzioni e dei sistemi di archiviazione adottati per il salvataggio dei contenuti digitali dell'Ateneo. Nel corso dello stesso anno si è deciso di intervenire avviando un progetto per la rea-lizzazione di un sistema centralizzato per l'archiviazione a lungo termine di tutti gli oggetti digitali prodotti dall'Ateneo. Il progetto, tuttora in corso, verrà realizzato tenendo conto di vincoli di sostenibilità economica e gestionale oltre che di scalabilità. Verranno inoltre definite policy, linee guida e altri aspetti organizzativi come ruoli, compiti ecc. dei vari attori coinvolti oltre che atti-vato un continuo colloquio con la community degli utilizzatori del sistema. Particolare attenzione verrà data alle caratteristiche di interoperabilità, in particolare con la rete della ricerca europea Eudat.

Keywords. Long Term Preservation, research data, dataset, iRODS, Eudat

Introduzione

Il progetto Arkive riguarda la realizzazione di un archivio per la conservazione, la gestione e diffusione degli oggetti digitali prodotti durante le attività di didattica, ri-cerca, divulgazione e amministrazione dell'Università degli Studi di Milano. Si tratta di un progetto strategico molto "sfidante" per la eterogeneità degli oggetti digitali che ci si propone di gestire e conservare. La progettazione del sistema sta impegnan-do da diversi mesi un gruppo di lavoro in cui sono presenti competenze trasversali riguardanti sia la gestione di rete e della sicurezza che la gestione di sistemi, la ge-stione dei dati della ricerca e gli archivi digitali: è necessario infatti affrontare il pro-blema tenendo conto contemporaneamente degli aspetti tecnologici, ma anche di quelli riguardanti la costruzione degli archivi digitali (dall'organizzazione della co-noscenza agli aspetti organizzativi) e delle particolarità dei dati della ricerca. Il pro-getto è nato dalla consapevolezza del forte rischio di perdita dei contenuti in formato digitale che da almeno due decadi costituiscono il formato nativo della quasi totalità della produzione scientifica e culturale dell'Università.

Palermo, 18-20 maggio 2022 - ISBN: 978-88-946629-1-7 - Creative Commons CC-BY

1. Conservare a lungo termine i dataset

Non esistono molti studi sistematici riguardanti la misurazione di quanto sia l'entità della perdita dei contenuti digitali dovuta a fattori accidentali, a obsolescenza o a scarsa cura nell'archiviazione. Sono stati pubblicati tuttavia alcuni articoli in cui vengono riportati risultati parziali e fortemente dipendenti dal contesto in cui vengo-no realizzati: misurare questo fenomeno non è infatti complesso dal punto di vista concettuale, ma lo è per la difficoltà di recuperare informazioni su qualcosa che è andato perso e di cui spesso non restano quasi tracce. Si pensi ad esempio al caso dei siti web relativi a progetti di digitalizzazione di beni culturali che diventano inacces-sibili e di cui rimangono riferimenti in altri contesti che portano a broken link. Per i dati della ricerca, informazioni possono essere tratte misurando per quali articoli scientifici rimangono disponibili e accessibili i dataset, considerando che in genere la produzione scientifica è più persistente.
Un articolo [1] del 2014 mostra come la perdita di accesso ai dataset relativi a un articolo scientifico che sia stato pubblicato è pari mediamente al 17% per ogni anno a partire dalla data di pubblicazione dell'articolo. Il che significa che dopo pochi anni la probabilità di poter accedere ai dati grezzi di un esperimento dopo che è stato pub-blicato decresce drasticamente: secondo lo studio dopo 20 anni risultava inaccessibile l'85% dei dataset relativi ai 511 articoli oggetto dell'indagine.

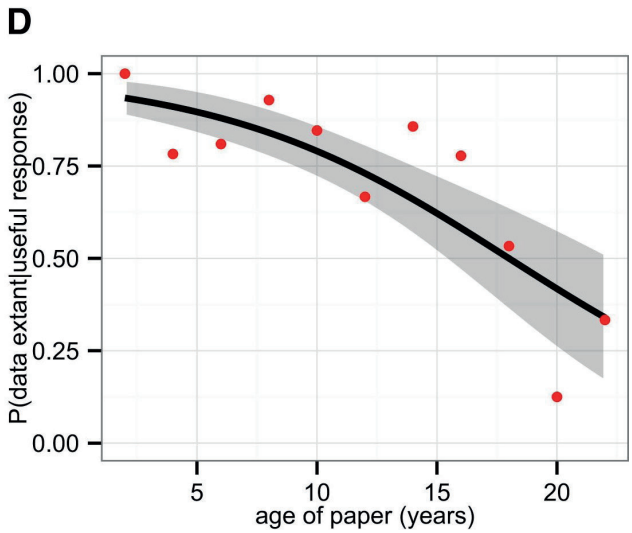

Fig. 1
Grafico del numero di dataset accessibili per anno di pubblicazione (tratto da [1])

Sono numeri quasi da "estinzione di massa" ed è impressionante la velocità con cui avviene questa perdita di dati. Il lavoro citato non è recente, sono passati 8 anni e viene da chiedersi se la situazione oggi sia migliorata, ma la risposta è che purtroppo lo stato attuale non è molto dissimile.
Uno dei mezzi con cui molti editori (Springer ed Elsevier ad esempio) cercano di arginare il problema della perdita di accesso ai dataset è la richiesta agli autori di sot-toscrivere una availability statement: cioè una dichiarazione di disponibilità dei dati, con relativi riferi-

menti, che deve accompagnare l'articolo.
Uno studio [2] più recente, del 2021, il cui obiettivo era la verifica della veridici-tà delle dichiarazioni di accessibilità dei dati da parte degli autori degli articoli, ha verificato, dopo un controllo puntuale, che per 4101 preprint soltanto il 23,1% (911) avevano reso aperti i propri dati. La situazione non migliorava di molto per quanto riguarda gli articoli effettivamente pubblicati in seguito e dotati di statement: solo per 59 su 151, ovvero il 38%, i dataset erano realmente disponibili. Quindi sembra che neppure una dichiarazione di disponibilità dei dati possa essere garanzia suffi-ciente per la continuità di accesso ai dati sperimentali.
Uno strumento più cogente è la richiesta da parte di agenzie finanziatrici della ri-cerca e dell'Unione Europea che i dati siano FAIR (Findable, Accessible, Interope-rable, Reusabile), requisito che condiziona molto spesso l'accesso ai finanziamenti. Un dataset è FAIR solo se ne viene assicurata la conservazione nel lungo periodo. E' chiaro però che per un autore rendere disponibili i propri dati in questa modalità è possibile solo a fronte della disponibilità di infrastrutture che lo permettano e queste infrastrutture devono essere affidabili, durevoli e facilmente utilizzabili. E se questo può essere già una realtà per alcuni domini di ricerca, relativi alla big science, dove esistono da tempo repositories disciplinari importanti (es. genomica) accade che per molti laboratori che svolgono studi su piccola e media scala in aree più specializzate non ci sia la possibilità di accedere a tali sistemi. In questi casi i dati rischiano di ri-manere conservati solo nei laboratori che li hanno prodotti e a volte di venire persi quando i membri del progetto di ricerca vanno via [3].
Si tratta quindi di un rischio di perdita di contenuti importante, e oltretutto non riguarda solo i dataset della ricerca, ma affligge tutto quanto viene prodotto in digi-tale e qui si può anche allargare il campo ai beni culturali che aggiungono altre speci-ficità al problema.
L'obiettivo di valorizzare ed estendere la fruizione di beni culturali (museali, li-brari, archivistici) ha dato origine a campagne di digitalizzazione che dai primi anni 2000 ha prodotto una notevole quantità di oggetti digitali. In questo caso molti pro-blemi sono causati dal fatto che la digitalizzazione non è stata quasi mai progettata per un uso nel lungo termine, ma per lo sviluppo di progetti "puntuali" di valorizza-zione (portali, siti web, cataloghi online) che a causa di problemi di obsolescenza spesso hanno cessato di essere fruibili dopo poco tempo. Non ci si è posti all'inizio il problema di che fine avrebbero fatto quei dati dopo qualche anno, se sarebbero stati ancora leggibili, comprensibili e fruibili. Spesso è presente una catalogazione del ma-teriale analogico di partenza, ma la metadatazione degli oggetti digitali è parziale o del tutto assente e il formato dei dati è stato scelto senza tenere conto delle caratteri-stiche utili per la conservazione.
Per quanto riguarda i dati della ricerca, l'Università degli Studi di Milano ha ne-gli scorsi anni avviato diverse iniziative a favore della scienza aperta e in particolare per la pubblicazione dei dati in modalità FAIR con l'adozione della piattaforma Sci-verse, tuttavia la componente di conservazione a lungo termine era ancora assente nell'Ateneo milanese.

2. Dunque, che fare?

In Università di Milano si è deciso di provare ad affrontare questo problema rea-lizzando

una infrastruttura “agnostica” o “olistica” per tutti gli oggetti digitali, cer-cando una soluzione per il problema visto nel suo complesso, in maniera unitaria.
Si è innanzitutto partiti con un’indagine conoscitiva, un questionario che è stato somministrato a tutti i potenziali produttori di oggetti digitali nell’ambito della ricer-ca e della didattica, vale a dire i dipartimenti, e per loro tramite i centri di ricerca la-boratori e così via. Sono stati coinvolti anche archivi, biblioteche e musei relativa-mente alle attività per le quali producono oggetti digitali, ovvero digitalizzazione di fondi librari e archivistici e digitalizzazione di oggetti museali.
Il risultato del questionario, in sintesi, è stato che l’esigenza di uno spazio sicuro su cui riversare i propri dati è ovunque sentita con urgenza da tutti i soggetti coinvol-ti. E’ stato stimato che pregresso e corrente richiedono 60PB di spazio dati per l’archiviazione, con un incremento annuale, per i prossimi 5 anni, di 20PB. E’ stato poi confermato che vengono utilizzati innumerevoli tipi di formati in parte proprieta-ri, l’uso di standard è più diffuso per i beni librari e archivistici mentre per i dati del-la ricerca e i beni culturali è molto limitato.
Dopo questa ricognizione si è cercato di individuare quali tecnologie erano dispo-nibili e già utilizzate in progetti analoghi di altre istituzioni. Le tecnologie dovevano avere le seguenti caratteristiche:

- gestire grandi quantità di storage, permettere una facile scalabilità del siste-ma e avere ottime caratteristiche di affidabilità e semplicità di gestione e di uso
- permettere la massima automazione delle operazioni di cura dei dati
- favorire la condivisione dei dati tra gruppi di ricerca entro e fuori UNIMI
- permettere un facile utilizzo del sistema da parte dei produttori dei dati
- permettere il riuso e la diffusione (regolata da policy di accesso) degli ogget-ti digitali

In altre parole si è cercato di individuare soluzioni che permettessero di attuare una strategia di long-term preservation in modo sostenibile per l’Ateneo. Si è quindi proceduto a realizzare:

• Un’infrastruttura di rete mista, a indirizzamento in parte pubblico e in parte privato, pensata per attuare la separazione dei flussi di dati (un indirizzamen-to per il monitoraggio e gestione diverso da quello per l’accesso degli utenti, da e da fuori rete di Ateneo, e diverso ancora per la comunicazione tra le macchine) a vantaggio sia della gestione che della sicurezza informatica.
• Un’infrastruttura di server virtuali: un pool di server virtuali che lavorano in parallelo ad alta affidabilità, anche in questo caso con caratteristiche di scala-bilità e semplicità (molti server piccoli e dedicati ciascuno ad un servizio)

È stato inoltre individuato il framework open source iRODS (https://irods.org/) per implementare le procedure di data curation e automatizzarle, con questa tecnolo-gia si è realizzata quindi una prima piattaforma di test. Per l’interoperabilità si è scel-to di adot-

tare i moduli di Eudat (https://eudat.eu/) che permettono l' integrazio-ne/federazione nella Collaborative Data Infrastructure (CDI) europea, supportando così la partecipazione dell'Ateneo ad EOSC (European Open Science Cloud https://eosc-portal.eu/about/eosc). Il pregio della tecnologia iRODS è l'estrema flessibilità nel poter definire in maniera astratta gli obiettivi dell'archivio, il set delle proprietà rilevanti gestite da management policies implementate attraverso procedures composte da rules e microservices organizzati in workflow. iRODS permette, una volta definita la politica di data curation, di implementarla semplicemente componendo workflow con i concetti e gli strumenti sopra elencati.

Al momento è ancora in corso la valutazione di quali tecnologie e soluzioni di storage, open source e commerciali, presentino le caratteristiche adatte (Isilon, Ceph, S3, Spectrum IBM ...), ma l'adozione della tecnologia iRODS, che permette di virtualizzare il file-system, consente modifiche della piattaforma storage senza do-ver alterare l'architettura del sistema.
Una volta definiti gli aspetti tecnologici si è tuttavia solamente a metà dell'opera dovendo poi coinvolgere le comunità produttrici di contenuti nella definizione delle policy di gestione dell'archivio e della politica di long term preservation. Dovranno essere fatte scelte condivise riguardanti la produzione e la gestione degli oggetti digi-tali (incluso indicazioni sulla scelta dei formati, la metadatazione e licenze d'uso) e, in ultimo, ma non meno importante, definire l'apparato organizzativo che avrà una componente centrale (legata alla gestione tecnologica) e una decentrata presso i Di-partimenti (legata alla gestione dei dati).

3. Conclusioni

In un Ateneo, in particolare se di dimensioni medio grandi, una strategia e una poli-tica per la conservazione centralizzata a lungo termine degli oggetti digitali sono una necessità sempre più urgente, ma non attuabile con soluzioni tecnologiche prèt-a-porter di solo storage. Sono tuttavia disponibili oggi alcune tecnologie e piattaforme applicative open source che semplificano la costruzione di sistemi di archiviazione orientati alla cura e manutenzione dei dati e che hanno caratteristiche di scalabilità e astrazione tali da rendere possibile la sostenibilità nel tempo. In particolare Integrated Rule-Oriented Data System (iRODS) permetterà di costruire una piattaforma di Ate-neo, ad alto grado di automazione dei processi, integrata con la tecnologia di storage che via via sarà selezionata in base alla scalabilità ed economicità e con interfacce utente semplici per favorire il deposito dei dati, la loro condivisione, riuso e pubbli-cazione da parte dei gruppi di ricerca.

Un ringraziamento particolare va al prof. Goffredo Haus, promotore di questo progetto, e ai colleghi del gruppo di lavoro Giorgio Bagnato, Matteo Zoppi e Loredana Rollandi che con la loro competenza ed entusiasmo lo stanno rendendo possibile.

Riferimenti bibliografici

1. Vines, Albert, et al. (2014) The Availability of Research Data Declines Rapidly with

Article Age. Current Biology, 1, Vol. 24 (https://doi.org/10.1016/j.cub.2013.11.014)
2. McGuinness LA, Sheppard AL (2021) A descriptive analysis of the data availability statements accompanying medRxiv preprints and a compari-son with their published counterparts. PLoS ONE 16(5)
(https://doi.org/10.1371/journal.pone.0250887)
3. Michael Eisenstein (2022) In pursuit of data immortality.
Nature 4 April 2022
(https://www.nature.com/articles/d41586-022-00929-3)

[1]https://www.springernature.com/gp/authors/research-data-policy/data-availability-statements/

[2]https://www.elsevier.com/authors/tools-and-resources/research-data/data-statement

Autori

Federica Zanardini federica.zanardini@unimi.it

Federica Zanardini si laurea in Fisica dello Stato Solido a Pavia, si specializza in Scienze dei Materiali e si occupa di fisica delle superfici fino al 1999. Nell'anno 2000 inizia a occuparsi di tecnologie informatiche e si trasferisce all'Università di Milano per dedicarsi alla creazione della Biblioteca Digitale dell'Ateneo che sviluppa e coordina fino al 2021. Nel corso del 2021 è entrata a far parte del gruppo di progetto per la realizzazione dell'infrastruttura di archiviazione a lungo termine dei dati della ricerca.
Dall'anno 2020 è docente del corso "Digitalizzazione, Digital Preservation e Digital Curation" nell'ambito del Master in Digital Humanities dell'Università degli Studi di Milano.