Metodologie computazionali per l'organizzazione di archivi nati digitalmente

Mariangela Giglio
[1] Università di Bologna, Italia – mariangela.giglio2@unibo.it

ABSTRACT (ITALIANO)
Il progetto si colloca nell'ambito della digital forensics e dell'archivistica, affrontando la complessità dell'organizzazione e attraversamento di archivi born digital attraverso un caso di studio specifico: l'analisi comparativa di 58 copie forensi derivate da floppy disk d'autore provenienti dall'Archivio Franco Fortini. L'obiettivo primario è l'elaborazione di una metodologia replicabile di organizzazione archivistica, allo scopo di fornire una classificazione deterministica e verificabile di materiali nativi digitali. A tal fine si propone un approccio sistematico basato sull'hashing dei materiali testuali contenuti nei floppy come indizio di vicinanza tra i dati. Tali hash diventano la base di una matrice binaria che permette di evidenziare le relazioni gerarchiche tra le diverse cartelle e, di conseguenza, di sistematizzare l'archivio e la definizione dei rapporti tra i diversi materiali.
Parole chiave: archivi born digital; clusterizzazione; preservazione digitale; organizzazione archivistica

ABSTRACT (ENGLISH)
Computational Methodologies for the Organization of Born-Digital Archives.
This project lies at the intersection of digital forensics and archival science, addressing the complexity of organizing and analyzing born digital archives through a specific case study: the comparative analysis of 58 forensic copies derived from the floppy disks of the Franco Fortini Archive. The primary objective is the development of a replicable methodology for archival organization, aiming to provide a deterministic and verifiable classification of born digital materials. The project adopts a systematic approach based on hashing the textual materials within the archive as an indicator of proximity or distance between various files. From this, a binary matrix was developed to highlight hierarchical relationships between different folders and, consequently, to systematize the archive and define the relationships between its various materials**.**
Keywords: born digital archives; clustering; digital preservation; archival organization

1. INTRODUZIONE

Il progetto si propone di esplorare e sistematizzare i materiali digitali dell'Archivio Franco Fortini, un archivio ibrido che contiene, in aggiunta a documenti cartacei, anche floppy disk oggi conservati presso la Biblioteca di Area Umanistica dell'Università di Siena. Di questi supporti, molte sono copie di backup desunte dall'ultimo hard disk dell'autore, copie trasposte successivamente in immagini forensi. Si tratta di un caso esemplare di complessità archivistica: sono presenti, infatti, almeno tre gruppi di copie di backup, realizzati in momenti diversi da operatori distinti, unitamente a materiali non d'autore e a una serie di copie parziali effettuate più recentemente da studiosi del Centro Fortini (Marrucci & Tinacci, 2005).
L'obiettivo di questo studio è duplice: sviluppare una metodologia replicabile per la gestione e la classificazione di archivi born digital complessi e, al contempo, applicare tale metodologia al caso specifico dell'Archivio Fortini, particolarmente emblematico per la sua complessità e per la stratificazione dei materiali conservati.
La gestione di archivi born digital, come è noto, presenta diverse complessità (Carbé, 2023; Carroll et al., 2011; Digital Preservation Coalition, 2015; Guercio, 2013; Marrucci & Tinacci, 2005), ulteriormente aggravate dalla limitata attenzione che si è data in passato alla corretta preservazione di questa tipologia di materiale, destinata a rapida obsolescenza. Il caso dei floppy conservati nel fondo Fortini non è privo di problematiche: manca una numerazione d'archivio e un ordine sistematico dei materiali, i gruppi di floppy non sono chiaramente distinti e le etichettature sono spesso incoerenti o del tutto assenti. L'archivio inoltre è caratterizzato da ridondanze e difformità strutturali di vario genere. Si è reso dunque necessario stabilire

in via preliminare un ordinamento adeguato, che non può prescindere dall'individuazione dei singoli gruppi di backup.
In questa direzione si è elaborato uno script che, basandosi sull'hashing dei contenuti testuali, possa garantire una classificazione dei materiali deterministica e replicabile. Questo approccio mira ad affrontare in maniera strutturata la complessità intrinseca dell'archivio, automatizzando le operazioni necessarie alla codifica delle relazioni tra i file. I risultati preliminari, sebbene per ora applicati a un singolo archivio, appaiono promettenti e lasciano intravedere una possibile estensione della metodologia ad altri archivi.[1] In questo senso il lavoro vorrebbe proporre una traccia metodologica che fornisca un modello replicabile per la classificazione e l'organizzazione di archivi born digital, offrendo uno strumento di facile uso per contesti archivistici analoghi.

2. METODOLOGIA

Nella gestione degli archivi born digital, le sfide derivate dalla loro complessità intrinseca e dalla varietà dei formati dei dati richiedono un approccio metodico e sistematico. A tale scopo, è stato sviluppato uno script Python dotato di un'interfaccia grafica intuitiva, concepito per facilitare l'adozione dello strumento anche da parte di utenti non specialisti.[2] Il design dello script privilegia la modularità, offrendo agli utenti la possibilità di adattare l'utilizzo dello strumento alle esigenze specifiche del proprio ambito archivistico. Questa struttura modulare permette di isolare e gestire individualmente ciascuna fase del processo, migliorando così la manutenibilità e la scalabilità del sistema.
L'architettura dello script si articola in sei fasi principali, ognuna corrispondente a una funzione chiave, progettata per trattare le problematiche legate alla complessità e alla ridondanza dei materiali. Ogni fase è concepita come un passaggio deterministico nella classificazione e nell'analisi comparativa dei contenuti, assicurando trasparenza e replicabilità dell'intero processo.

3. LO SCRIPT: FUNZIONI CHIAVE

Fase preliminare: preparazione dell'ambiente di lavoro

Il lavoro preliminare consiste nella configurazione di un ambiente di lavoro separato dai file originali, un passaggio indispensabile per assicurare una manipolazione sicura e controllata dei dati. In questa fase, i documenti originali vengono trasferiti in una nuova directory di lavoro denominata Working Directory (WD), che replica esattamente la struttura originaria dei dati, mantenendo inalterati i metadati. Nel contesto dell'archivistica digitale è imperativo lavorare su copie dei file originali per mitigare il rischio di danneggiamenti o alterazioni accidentali. Questa metodologia, ampiamente riconosciuta e attestata (Chassanoff et al., 2016; Ries, 2022), è diventata una prassi per salvaguardare l'integrità del materiale originale durante l'analisi e la manipolazione dei dati. L'utilizzo di copie assicura infatti che eventuali errori di elaborazione o fenomeni di corruzione del materiale incidano unicamente sulle versioni duplicate, preservando così l'integrità del documento originario (Carroll et al., 2011).

Fase 1: Estrazione di file legacy

La fase iniziale dello script è preminentemente volta all'estrazione dei contenuti testuali dai documenti del corpus, operazione imprescindibile per archivi contenenti file in formati legacy e/o materiali non rilevanti ai fini di analisi, quali file di sistema, residui di backup e metadati vari. La presenza di file legacy con estensioni non standard rappresenta una sfida comune negli archivi ibridi o born digital (Light, 2010). Per affrontare questa problematica – limitatamente ai file testuali – è stato adottato l'uso di LibreOffice, software che incorpora avanzate librerie dedicate alla gestione di file testuali obsoleti.[3]
Al fine di garantire un'esclusione mirata e accurata, sono stati automaticamente omessi elementi considerati non essenziali, quali file temporanei e nascosti (e.g., .DS_Store e desktop.ini), nonché directory di sistema (e.g., resource.frk e aree non allocate), consentendo inoltre all'utente l'aggiunta di filtri personalizzati per l'esclusione. I file ritenuti pertinenti sono stati successivamente convertiti in formato

---

[1] Si pensi, a titolo d'esempio, al fondo di Francesco Pecoraro conservato al Centro Manoscritti di Pavia, al computer del filosofo e matematico Imre Toth conservato presso la Biblioteca di Area Umanistica di Siena o, ancora, ai materiali digitali conservati all'Archivio Bonsanti del Gabinetto Vieusseux (Carbé, 2023, p. 41,92).

[2] Lo script (in forma anonimizzata) è disponibile su Github.

[3] Per ulteriori informazioni sulle librerie incluse in LibreOffice è possibile fare riferimento alla pagina Licensing and Legal information del software (cons. 01/01/2025).

.odt mediante l'interfaccia a riga di comando (Command Line Interface - CLI) di LibreOffice. Durante questa fase ci si mantenuta inalterata la struttura gerarchica originale dell'archivio, una pratica che assicura che ciascun elemento del corpus possa essere precisamente ricollocato nella sua posizione originaria, facilitando l'analisi successiva e conservando l'integrità contestuale dei dati estratti. L'estrazione dei materiali testuali in formati standard garantisce che la fase di hashing successiva possa avvenire senza essere influenzata da fattori non significativi quali i metadati dei file o contenuti non rilevanti.

Fase 2: Hashing

La prosecuzione dello script introduce una fase di hashing che costituisce un momento cruciale della clusterizzazione, influenzando direttamente l'efficacia delle fasi processuali successive. L'hashing è un processo computazionale che trasforma un insieme di dati di dimensioni variabili in una rappresentazione univoca, denominata hash, una sorta di "impronta digitale" di ciascun file. Questa operazione permettere confronti rapidi tra file, ed è utile sia per la verifica dell'integrità dei dati sia per la loro autenticazione (Chi & Zhu, 2018; Roussev, 2009).

Al fine di ottenere una rappresentazione fedele dei documenti, la tecnica implementata esclude metadati, percorsi dei file e altre variabili esterne che potrebbero alterare l'hash generato, focalizzandosi sul solo contenuto testuale estratto dai file.

Utilizzando l'algoritmo SHA-256[4] ogni documento in formato .odt è stato processato per ottenere un hash univoco, garantendo così una precisa identificazione dei file anche in presenza di minime variazioni. Questi hash sono stati successivamente catalogati in un dizionario in cui ogni chiave rappresenta una cartella, mentre i valori sono gli elenchi degli hash dei file contenuti nella directory stessa e nelle sue sottodirectory. Questo metodo consente di ridurre la complessità del dataset, focalizzandosi sulla distribuzione dei contenuti testuali e facilitando la comparazione dei file all'interno dell'archivio.

Fase 3: Creazione di una matrice binaria hash/directory

La terza fase del processo consiste in una conversione del dizionario degli hash in una matrice binaria che stabilisce una correlazione tra ogni hash univoco alle directory che lo contengono, offrendo così una rappresentazione formale della presenza o assenza di ciascun hash unico nelle diverse directory. Tale struttura costituisce una prima schematizzazione che delinea le relazioni tra i contenuti e le loro posizioni all'interno dell'archivio.

Nella matrice, ogni riga è associata a un hash distinto, mentre ciascuna colonna corrisponde a una determinata cartella. Le celle sono riempite con valori booleani, zero o uno, indicanti rispettivamente l'assenza o la presenza dell'hash specifico nella directory corrispondente. Questa rappresentazione strutturale facilita l'identificazione rapida delle relazioni tra i file distribuiti nei vari supporti di archiviazione, permettendo di evidenziare le copie identiche attraverso colonne perfettamente allineate e di rilevare le divergenze, parziali o totali, mediante colonne con pattern dissonanti.

Il numero di righe nella matrice è pari alla quantità di hash unici prodotti, mentre il numero di colonne riflette il numero di directory di livello superiore esaminate.

	Cartella α	***Cartella β***	***Cartella γ***	***Cartella δ***	***Cartella ε***
Hash 1	1	0	1	1	0
Hash 2	0	0	0	1	1
Hash 3	1	1	0	0	0
Hash 4	1	0	1	1	1
Hash 5	1	0	1	1	0

Tab.1 Esempio di output in matrice binaria

Fase 4: Matrice di similarità

Nella fase successiva del processo, basandosi sulla matrice binaria sviluppata precedentemente, si procede alla configurazione di una matrice di similarità che rappresenti formalmente il grado di affinità tra le varie

[4] SHA-256 è una funzione di hash crittografica ampiamente impiegata per garantire l'integrità, l'autenticità e la sicurezza dei dati in varie applicazioni, incluse le firme digitali, i codici di autenticazione dei messaggi e la tecnologia blockchain (Devi & Jayasri, 2023; Gilbert & Handschuh, 2004).

directory. La matrice, di natura simmetrica, attribuisce un valore diagonale costante pari a 1, segnalando l'identità di ciascuna directory rispetto a sé stessa. Le presenze o le assenze di hash univoci, codificati precedentemente nella matrice binaria, vengono così trasformati in valori numerici che riflettono la similarità tra i contenuti delle directory. All'interno di questa matrice, ogni riga corrisponde a un hash distinto, mentre ogni colonna rappresenta una specifica cartella. Le celle, popolate da valori booleani (0 o 1), indicano rispettivamente l'assenza o la presenza di un determinato hash nella cartella corrispondente. Questa configurazione strutturale agevola l'identificazione immediata delle relazioni tra i supporti di memorizzazione, permettendo di evidenziare sia le copie esatte – attraverso colonne identiche – sia le divergenze parziali o totali manifestate attraverso colonne con pattern divergenti.
Il principio su cui si basa la matrice di similarità deriva dal confronto tra le colonne della matrice binaria, dove ogni colonna simbolizza una directory e ogni riga un hash unico. La misura di similarità tra due colonne viene calcolata attraverso l'identificazione delle discrepanze binarie, ovvero le righe in cui un hash appare in una colonna ma è assente nell'altra. Questa analisi produce risultati che variano su una scala di similarità: un valore di 1 indica una sovrapposizione totale, mentre un valore di 0 denota una divergenza completa.
La formula utilizzata per calcolare la similarità è:

$$S = 1 - \frac{d}{nmax}$$

Form.1 Formula per calcolare la similarità

Dove S rappresenta il valore di similarità, d indica il numero di differenze binarie tra due colonne, e $nmax$ è il numero massimo di hash presenti in una singola cartella.
Questa formula viene implementata nella funzione per attribuire un punteggio di similarità variabile da 0 a 1 a ogni coppia di cartelle esaminata.
Consideriamo adesso il caso di due cartelle, indicate come α e β, per illustrare l'applicazione pratica della formula di similarità.

$$S\alpha\beta = 1 - \frac{d\alpha\beta}{nmax}$$

Form.2 Formula per calcolare la similarità tra le cartelle α e β

Il valore di similarità tra α e β può dare diversi risultati a seconda della effettiva somiglianza tra le due cartelle. A titolo di esempio:

- $S\alpha\beta=1$ indica che α e β sono cartelle identiche
- $S\alpha\beta=0$ indica che α e β sono cartelle totalmente diverse (non condividono nessun file)
- $0.5< S\alpha\beta<1$ indica che α e β presentano delle sovrapposizioni parziali
- $S\alpha\beta=0.01$ indica che α e β sono profondamente diversi ma condividono alcuni file
- $S\alpha\beta=0.8$ indica la presenza di numerose sovrapposizioni

Nel contesto dell'Archivio Franco Fortini, questa distinzione è particolarmente rilevante: ogni gruppo di backup è composto da floppy che contengono dati differenti, ma la presenza di file identici tra gruppi diversi produce corrispondenze puntuali all'interno della matrice. Di conseguenza, punteggi pari a 1 non indicano necessariamente appartenenza a uno stesso gruppo di backup, bensì replicazione esatta tra floppy gemelli. Al contrario, la struttura globale della matrice di distanza, con cluster di somiglianza parziale, consente di identificare i gruppi di backup e comprendere le relazioni tra essi. Questa fase rappresenta il punto di arrivo del processo analitico, traducendo i dati grezzi in una visione strutturata e quantitativa delle relazioni tra i materiali.

Fase 5: Visualizzazione dei risultati
La fase conclusiva dello script si avvale della matrice di similarità elaborata nelle fasi antecedenti per generare visualizzazioni che delineano chiaramente le relazioni tra i documenti. Per implementare il clustering gerarchico, essenziale per queste visualizzazioni, la matrice di similarità è trasformata in una matrice di distanza, definendo la distanza come $1-S$.[5] Questa trasformazione è indispensabile perché

[5] Similarità, vedi Formula 1.

numerose tecniche di clustering operano più efficacemente con le metriche di distanza piuttosto che di similarità. Si è tuttavia scelto di conservare la matrice di similarità come output principale, in quanto offre una rappresentazione intuitiva e facilmente accessibile, sia per l'interpretazione umana sia per l'elaborazione informatica, e può essere convertita all'occorrenza per analisi ulteriori.
I dendrogrammi vengono generati impiegando la funzione "linkage" di SciPy, partendo dalla matrice di distanza condensata, derivata dalla matrice di distanza quadrata[6] elaborata in precedenza. Metodo di linkage scelto per l'analisi è quello di Ward, noto per la sua efficacia nel minimizzare la varianza all'interno dei cluster. Questo approccio assicura un raggruppamento dei dati coerente e informativo, come evidenziato da Großwendt et al. (2019) e Strauss & Von Maltitz (2017). Per consentire una maggiore flessibilità è stata comunque lasciata all'utente la possibilità di personalizzare metodo di linkage e distanza.

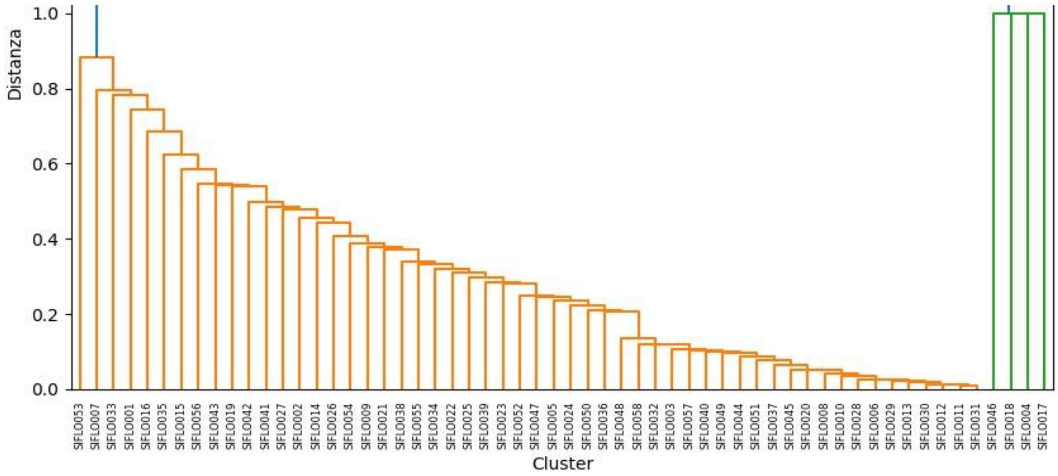

Fig.1 Il dendrogramma rappresentante il raggruppamento dei materiali

4. VALIDAZIONE DEI RISULTATI

Per valutare l'attendibilità dei risultati ottenuti tramite la procedura di clusterizzazione automatizzata, è stata condotta un'operazione di validazione manuale che ha assunto il ruolo di *ground truth* dell'intero processo. Tale attività ha previsto un'analisi approfondita dei contenuti testuali dei file, della struttura delle directory presenti nei supporti e dei metadati espliciti (come etichette, nomi di cartelle e formati di file), con l'obiettivo di individuare somiglianze ricorrenti e relazioni funzionali all'interno del corpus. A questo si è affiancata un'analisi forense condotta direttamente sulle immagini dei floppy disk, con particolare attenzione ai *timestamp* di creazione e modifica, ai nomi utente e ad altre tracce residue conservate nei file di sistema. Tali metadati nascosti, spesso non rilevabili tramite semplice ispezione documentale, hanno permesso di formulare ipotesi di raggruppamento indipendenti dal contenuto, fornendo ulteriori indizi di prossimità tra i supporti.
Il confronto tra la clusterizzazione *human validated* e i risultati prodotti dallo script ha evidenziato una sostanziale coerenza tra i due modelli. I principali gruppi di backup sono stati riconosciuti correttamente dall'algoritmo, e i dendrogrammi generati riflettono con buona approssimazione le relazioni gerarchiche ipotizzate manualmente. Le copie identiche e le somiglianze parziali rilevate automaticamente trovano corrispondenza nella struttura fisica e logica degli archivi digitali, a conferma della solidità dell'approccio computazionale. Le discrepanze riscontrate si concentrano su casi ambigui o liminali, in cui la collocazione archivistica risulta influenzata da fattori extratestuali, come la frammentazione di una directory su più supporti o operazioni successive di duplicazione.
Per consolidare ulteriormente la validità del metodo, è stata condotta una validazione quantitativa sulla porzione di corpus clusterizzata manualmente. Le performance della clusterizzazione sono state misurate

[6] La matrice quadrata, usata per rappresentare le distanze tra coppie di punti in una tabella bidimensionale, è caratterizzata dalla simmetria (ad esempio, la distanza tra i punti a e b è uguale a quella tra b e a) e da elementi diagonali nulli. Una versione più compatta è la matrice condensed, che si presenta come un array monodimensionale. Quest'ultima include solo le distanze sopra o sotto la diagonale principale, eliminando ridondanze e valori zero della diagonale, il che riduce significativamente l'uso della memoria e aumenta l'efficienza computazionale, soprattutto per matrici di grandi dimensioni.

attraverso un insieme di metriche standard per la valutazione del clustering non supervisionato, tra cui l'Adjusted Rand Index (ARI), la Jaccard macro-average, e le misure derivate dalla mutual information: Homogeneity, Completeness e V-Measure.[7] I risultati confermano una buona coerenza strutturale tra i due modelli (ARI ≈ 0.69), anche se la frammentazione di alcuni gruppi e la parziale sovrapposizione evidenziata dallo Jaccard score (≈ 0.46) indicano margini di miglioramento.
Le limitazioni principali emerse riguardano la generalizzabilità del metodo a contesti archivistici diversi. L'efficacia della procedura dipende infatti in larga misura dalla qualità e coerenza dei dati disponibili: archivi con file non testuali, strutture di file system danneggiate o contenuti altamente eterogenei possono ridurre l'accuratezza dei risultati. Inoltre, l'uso di tecniche di hashing e tokenizzazione privilegia una logica di somiglianza contenutistica, che può non cogliere appieno l'organizzazione semantica o funzionale di alcuni archivi. Ciò risulta particolarmente evidente in casi in cui directory affini sul piano concettuale differiscano sensibilmente nella forma o nel contenuto. La strategia adottata si dimostra quindi solida in contesti coerenti, ma richiede un affinamento metodologico per gestire anomalie locali, supporti parziali o materiali liminali, e per una eventuale estensione del modello a dataset meno strutturati.

5. CONCLUSIONI

Le clusterizzazioni effettuate dallo script sono state messe a confronto con le ipotesi preliminari di raggruppamento basate sull'osservazione diretta dei file, evidenziando risultati non ancora perfetti ma promettenti e significativi per ulteriori indagini. Le precedenti analisi dei materiali digitali dell'archivio non avevano permesso di identificare tutte le relazioni significative, come ad esempio le somiglianze parziali tra diverse cartelle, che sono invece emerse chiaramente nel raggruppamento elaborato con lo script. L'esame del dendrogramma ha rivelato la presenza di ulteriori copie correlate ai tre gruppi principali di backup (i floppy 046, 018, 004, 017 evidenziati in verde nella Figura 1). Questi raggruppamenti, sebbene individuabili attraverso un'analisi tradizionale, potrebbero non essere riconosciuti senza un esame completo dell'archivio: risalta l'efficacia del raggruppamento in insiemi di tre, suggerito dall'algoritmo, che pare riflettere accuratamente la suddivisione originaria in tre gruppi di backup. Permane, tuttavia, la presenza di alcune clusterizzazioni problematiche o errori di riconoscimento che potrebbe essere emendata con ulteriori affinamenti dell'algoritmo di clustering. Nel caso specifico del fondo Fortini questa interrelazione potrebbe, tuttavia, essere un residuo del processo di copia sequenziale in più floppy di dimensioni diverse, ipotesi che necessita di un'analisi più approfondita per essere supportata.
Mediante un approccio modulare e automatizzato, si è affrontata la complessità di un corpus caratterizzato da ridondanze, formati eterogenei e strutture non uniformi. L'adozione di tecniche computazionali, come l'hashing dei contenuti testuali e la costruzione di matrici binarie e di distanza, ha permesso di tradurre la complessità intrinseca dei dati in una rappresentazione strutturata e quantificabile. Le varie fasi del progetto sono state armonizzate all'interno di una pipeline che ambisce a fornire replicabilità e trasparenza, componenti essenziali per l'applicazione di tali metodologie in contesti archivistici analoghi. Le visualizzazioni grafiche, infine, hanno supportato l'interpretazione dei dati, rendendo leggibili le relazioni interne ai materiali. Il lavoro evidenzia come la combinazione di analisi statistiche automatiche e revisioni manuali possa contribuire alla migliore gestione di archivi born digital complessi. Sebbene il focus sia stato l'Archivio Franco Fortini, la metodologia proposta potrebbe essere estesa ad altre realtà archivistiche, ponendo le basi per un approccio più generalizzabile alla gestione di materiali nativi digitali.

RINGRAZIAMENTI

Desidero esprimere la mia sincera gratitudine alla professoressa Emmanuela Carbé per avermi indicato una preziosa traccia metodologica da seguire.
Ringrazio inoltre Francesco Garosi, esperto di tecnologie digitali, per il supporto e la competenza messi a disposizione nella realizzazione della componente informatica del lavoro.

[7] Le metriche di validazione includono: Adjusted Rand Index (che misura la concordanza tra assegnazioni indipendentemente dai nomi dei cluster), Jaccard macro-average (che valuta la sovrapposizione tra le classi), Homogeneity (quanto ciascun cluster contiene elementi di una sola classe), Completeness (quanto tutti gli elementi di una classe sono nello stesso cluster) e V-Measure (media armonica tra Homogeneity e Completeness).

BIBLIOGRAFIA

Carbé, E. (2023). *Digitale d'autore: Macchine, archivi e letterature*. Firenze University Press. https://doi.org/10.36253/979-12-215-0023-3.

Carroll, L., Farr, E., Hornsby, P., & Ranker, B. (2011). A Comprehensive Approach to Born-Digital Archives. *Archivaria*, *72*, 61–92.

Chassanoff, A., Woods, K., & Lee, C. A. (2016). Digital Preservation Metadata Practice for Disk Image Access. In A. Dappert, R. S. Guenther, & S. Peyrard (A c. Di), *Digital Preservation Metadata for Practitioners* (pp. 99–109). Springer International Publishing. https://doi.org/10.1007/978-3-31943763-7_8

Chi, L., & Zhu, X. (2018). Hashing Techniques: A Survey and Taxonomy. *ACM Computing Surveys*, *50*(1), 1–36. https://doi.org/10.1145/3047307

Devi, S. S., & Jayasri, K. (2023). Esha-256: An Enhanced Secure Cryptographic Hash Algorithm for Information Security. *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, 952–958. https://doi.org/10.1109/ICACRS58579.2023.10405239

Digital Preservation Coalition. (2015). *Digital Preservation Handbook. 2nd edition*. https://www.dpconline.org/handbook

Gilbert, H., & Handschuh, H. (2004). Security Analysis of SHA-256 and Sisters. In M. Matsui & R. J. Zuccherato (A c. Di), *Selected Areas in Cryptography* (Vol. 3006, pp. 175–193). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-24654-1_13

Großwendt, A., Röglin, H., & Schmidt, M. (2019). *Analysis of Ward's Method*. 2939–2957. https://doi.org/10.1137/1.9781611975482.182

Guercio, M. R. (2013). *Conservare il digitale: Principi, metodi e procedure per la conservazione a lungo termine di documenti digitali* (1. ed). Laterza.

Light, M. (2010). *Designing a Born-Digital Archive*. 1.

Marrucci, M., & Tinacci, V. (2005). L'edizione di uno scritto a testimonianza plurima, cartacea e informatica: Un giorno o l'altro di Franco Fortini. *Filologia italiana 2*.

Ries, T. (2022). Digital history and born-digital archives: The importance of forensic methods. *Journal of the British Academy*, *10*, 157–185. https://doi.org/10.5871/jba/010.157

Roussev, V. (2009). Hashing and Data Fingerprinting in Digital Forensics. *IEEE Security & Privacy Magazine*, *7*(2), 49–55. https://doi.org/10.1109/MSP.2009.40

Strauss, T., & Von Maltitz, M. J. (2017). Generalising Ward's Method for Use with Manhattan Distances. *PLOS ONE*, *12*(1), e0168288. https://doi.org/10.1371/journal.pone.0168288