Dai metadati ai dati, dai contesti ai contenuti: aumentare la descrizione archivistica

Federico Valacchi[1]
[1] Università degli Studi di Macerata, Italia – federico.valacchi@unimc.it

ABSTRACT
L'intervento si propone di valutare in chiave comparativa la dimensione culturale, tecnica e scientifica degli strumenti archivistici per individuare tecniche euristiche che agiscano sui contenuti e sui dati e vadano oltre i fisiologici limiti quantitativi della ricerca archivistica che si manifestano anche negli archivi digitali o digitalizzati

PAROLE CHIAVE
Descrizione archivistica; archivi digitali; collezioni digitali; learning machine; strumenti di ricerca

1. INTRODUZIONE

La descrizione archivistica, per lunga e inevitabile tradizione, procede muovendo dallo studio dei contesti verso un'individuazione necessariamente approssimativa dei contenuti[1]. La buona volontà della mediazione, nella maggior parte dei casi, si deve infatti confrontare con una brutale dimensione quantitativa. Gli archivi sono oceani di informazione e gli strumenti e gli approcci che potremmo definire tradizionali fanno oggettivamente fatica a dare istruzioni davvero puntuali ai naviganti. D'altra parte è altrettanto inevitabile mantenere i contenuti agganciati ai contesti, se vogliamo continuare a parlare di archivi in senso proprio, per quanto allargati.

Dal punto di vista squisitamente metodologico si pone quindi il problema di andare oltre il canone descrittivo senza tradire le coordinate di fondo della descrizione archivistica e del metodo stesso.

Tempus fugit e anche le nostre idee di descrizione e, soprattutto di fruizione, devono fare i conti con la voracità dell'obsolescenza innescata dalle violente e pervasive accelerazioni tecnologiche cui la nostra società nel suo insieme è sottoposta. In particolare, quello che potremmo definire un costume tecnologico indotto ha diversificato l'utenza degli archivi, rendendola più esigente[2]. Siamo ormai abituati ad ottenere risposte piuttosto che a porci domande. Ci muoviamo dentro alle logiche per certi aspetti perverse dei motori di ricerca. Molti utenti faticano perciò a comprendere come proprio gli archivi, luoghi deputati alla custodia e all'uso dell'informazione, stentino a rispondere in maniera puntuale alle loro interrogazioni. Come abbiamo detto, esistono ragioni incontestabili per spiegare questa approssimazione informativa, ma limitarsi a giustificarla non basta più.

"La domanda di risposte" non può più essere ignorata, anche alla luce di una tecnologia che amplifica i bisogni e sembra suggerire soluzioni allettanti.

Il problema è reale e complicato. Mette in gioco le tecnologie, anche venture, e le politiche di digitalizzazione, ma tira in ballo anche la nostra idea di mediazione. Le risposte che cerchiamo possono infatti incidere sulla descrizione archivistica, magari *aumentandola* per insegnare alla artificiale intelligenza delle macchine a districarsi tra le parole dei documenti.

Si potrebbe obiettare che c'è il rischio di snaturare la disciplina, ma in realtà l'archivistica insegue da sempre la mutevolezza degli archivi e dei bisogni che essi devono soddisfare. Si può imparare senza abdicare, ci si possono porre obiettivi ambiziosi senza rinunciare ai propri metodi e ai propri strumenti. Se

[1] Antonio Romiti, «I mezzi di corredo archivistici e i problemi dell'accesso», *Archivi per la storia* III, fasc. 2 (1990): 217–46.
[2] Barbara Lazenby Craig, «Old Myths in New Clothes: Expectations of Archives Users», *Archivaria* 45, (1998): 118–26.

il formato e la fisionomia dell'inventario archivistico potranno modificarsi e diventare qualcosa di diverso dalla percezione che ne abbiamo sempre avuto, non necessariamente verranno meno le ragioni profonde per cui si producono strumenti di ricerca.
Bisogna chiedersi se sia possibile aumentare la portata informativa degli strumenti archivistici in senso ampio, salvaguardando i sistemi di relazioni che governano ogni fondo archivistico, ma permettendo al tempo stesso agli utenti di estrarre più agevolmente il succo informativo dei singoli documenti, il cui reperimento giustifica l'intero lavoro archivistico.
Le tecnologie di cui disponiamo in misura apparentemente inesauribile possono darci indicazioni e perfino risposte a questo riguardo, ma dobbiamo essere disponibili a metterci in gioco, ferma restando la conoscenza di base di ogni fondo archivistico: l'archivio va prima descritto e poi riordinato, altrimenti ogni strategia empirica e tecnologica sarà vana se non controproducente.
Se l'archivio è ordinato, come abbiamo detto, potremmo cercare la soluzione in un concetto aumentato di descrizione archivistica, all'interno del quale ricondurre intanto processi di indicizzazione e/o di trascrizione selettiva dei documenti digitalizzati, finalizzata ad assecondare l'allenamento delle macchine nel riconoscimento automatico del testo. L'intento è quello di svegliare dal suo torpore l'immagine digitale, mummificata nel formato di copia fotografica dell'originale analogico.
Prima che tecnica e tecnologica la questione è metodologica e, in un certo senso, perfino antropologica. Occorre intanto rivisitare acquisizioni consolidate, per accettare soprattutto che *soggettazione* o *materie* possono non essere bestemmie archivistiche se le si declina nel modo opportuno e nella dovuta armonia tecnologica.
Sembra poi evidente che in questi ipotetici scenari l'archivistica non basta più a sé stessa. Si impongono nuove e costruttive alleanze con le altre discipline dell'universo documentario, allargando lo spettro della collaborazione anche ai diversi domini di gestione dell'informazione in senso stretto. Si va dalla paleografia alle digital humanities nel senso più ampio e nobile dell'espressione e si va oltre le pur sacrosante esigenze euristiche. Quella che si annuncia infatti è una battaglia della conoscenza contro la (dis)informazione digitale, che mette in gioco tutte le discipline di area LIS e, più in generale, suggerisce un confronto franco e concreto con le digital humanities, in cerca di comuni spazi di ulteriore sviluppo[3].

2. METODO, STRUMENTI E RICERCA

Ogni fondo archivistico si manifesta solo nel suo riuso informativo nel tempo. Sono le sollecitazioni esterne a determinarne il valore specifico e per questa ragione il processo di mediazione archivistica accompagna l'intero ciclo vitale e ad esso deve adeguarsi. Gli strumenti di ricerca, ed in particolare gli inventari, sono essi stessi prodotti in divenire e sono prima di tutto testimonianze peculiari del clima culturale e scientifico da cui scaturiscono. L'inventario, dal punto di vista dell'uso, è esso stesso parte del contesto che mira a ricostruire. La sua necessaria approssimazione lo apre paradossalmente a possibili implementazioni anche successive alla "pubblicazione", soprattutto quando ci si riferisca a banche dati di descrizioni archivistiche. La ricerca che muove dalla ricerca asseconda un'economia circolare delle informazioni che, almeno in parte, può supplire ai limiti fisiologici cui abbiamo accennato[4]. Gli strumenti non sono astrazioni euristiche. Sono piuttosto il complesso risultato di un sistema di costruzione della conoscenza che affonda le sue radici in un metodo che nella sua essenza continua a funzionare, anche perché oggettivamente è l'unico che abbiamo. L'insieme delle risorse sempre più raffinate che la mediazione archivistica ha reso disponibili nel tempo ha dato sicuramente risultati più che soddisfacenti. La riflessione mai interrotta sulla natura e le finalità della descrizione archivistica ne è la tangibile dimostrazione e garanzia. Il metodo è l'impianto sintattico e grammaticale della lingua con cui si esprimono gli archivi. E ci protegge dal rischio concreto di un'ingovernabile anarchia documentaria.

[3] Marilena Daquino e Francesca Tomasi, «Digital Humanities e Library and Information Science. Attraverso le lenti dell'organizzazione della conoscenza», *Bibliothecae.it* 5, fasc. 1 (2016): 130–50, https://doi.org/10.6092/ISSN.2283-9364/6109
[4] Stefano Gardini, «Economie circolari dell'archivio: la carte di utenti e studiosi come archivi derivati», *Nuovi Annali della Scuola speciale per archivisti e bibliotecari* XXXV (2021): 237–77.

Il metodo storico, "archivistico" per definizione, continua a fare il suo lavoro, assecondato dagli standard di descrizione di prima e seconda generazione, ISAD(G) e RiC in testa, su cui avremo modo di tornare brevemente più avanti. Il risultato finale dei processi descrittivi continua a manifestarsi in una gamma di strumenti di ricerca che in molti casi sono costretti a puntare al contesto più che al contenuto. La massiccia digitalizzazione delle risorse archivistiche nel suo insieme, di fatto, non ha permesso di superare questo fondamentale limite, dal momento che si è fin qui limitata a riproporre strumenti e metodi consolidati, rivisitandoli alla luce di una maggiore potenza di calcolo. Il risultato della ricerca *on demand*, per chiamarla così, continua ad essere più una speranza che una certezza.

Abbiamo il metodo e abbiamo gli strumenti ma né l'uno né gli altri sono incisi sulla pietra[5]. In ragione di quella duttilità che si richiamava sopra, il metodo e gli strumenti che ne scaturiscono funzionano quando assecondano le ragioni effettive della produzione e della fruizione. È vero che ad un livello squisitamente funzionale le tecniche e le tecnologie con cui si costruisce il sistema di mediazione ne possono modificare nel tempo la configurazione, migliorandone anche il rendimento. Sarebbe però fuorviante confidare in un'evoluzione di taglio banalmente tecnologico. Indipendentemente dalle sue potenzialità e dalle sue caratteristiche strutturali, ogni inventario ha sempre risposto innanzitutto a problemi di organizzazione e restituzione di particolari famiglie di metadati. Un inventario è, appunto, un sistema *strutturato* di dati sui dati e tale resta indipendentemente da come lo si costruisce e restituisce.

Il metodo quindi ci serve ancora, anche se quando ci affacciamo sul marcato polimorfismo contemporaneo, e in particolare sugli archivi digitali, qualcosa può cambiare.

Il cambiamento più marcato e gravido di conseguenze non è solamente di natura meccanica e legato ai mezzi di produzione e alla natura dei supporti. Tendono infatti a modificarsi soprattutto le logiche di produzione e uso da cui poi deriva la diversificazione dei mezzi e degli strumenti.

Per capire meglio conviene innanzitutto definire quale sia *l'archivio* di cui parliamo, precisando che qui ci riferiamo ad aggregazioni digitali, sia native che generate a partire da processi di acquisizione di fondi archivistici analogici. Si può allora introdurre intanto una distinzione di massima, per quanto grossolana, tra due fondamentali tipologie. Da un lato stanno gli archivi informatici, tra i quali per estensione si possono considerare anche i siti web, con i problemi conservativi che pongono, e dall'altro la mole crescente di archivi o di porzioni di archivio digitalizzati a partire da consolidate sedimentazioni analogiche.

Nel primo caso vanno innanzitutto segnalate trasformazioni che già incidono in profondità sull'auspicabile "storicizzazione" di questi complessi documentari, sia in termini di processi conservativi che di adeguata contestualizzazione. L'impatto metodologico è forte, perché si mette in discussione la rassicurante univocità del *creator.* Si modifica il flusso funzionale della produzione, che non scaturisce più in maniera univoca da un solo soggetto, magari fortemente strutturato. La filiera documentaria tende a diluirsi, inseguendo le esigenze e le lusinghe di un'interoperabilità che non è solo linguaggio di scambio tra le macchine ma modo di agire di buona parte dei soggetti produttori.

La stessa conservazione, poi, perde consistenza e tracciabilità nei meandri di una delocalizzazione fisica che distrugge l'idea stessa di policentrismo e pone problemi di percezione unitaria ed univoca delle *universitas rerum* documentarie. L'archivio c'è, ma non si vede, e governarlo correttamente è più complicato. Si pone insomma la questione di un effettivo approccio storico e culturale a questi archivi. Non basta più manifestare la volontà di difendere la memoria digitale dall'obsolescenza o limitarsi a pensare a strategie di sopravvivenza che garantiscano la *long time preservation* degli oggetti digitali, siano esse migrazioni, cloud o blockchain. Il passo da fare sembra essere quello di porre la questione al giusto livello politico e culturale, riflettendo seriamente su un modello conservativo nuovo e adeguato al presente e al futuro della produzione documentaria.

Capire *come* conservare è vitale ed è la nuova urgente configurazione dell'idea di base di tutela. Se è vero che si conserva per consultare, è urgente riflettere anche sul *perché*.

La conservazione di lungo periodo costruisce *archivi storici in senso proprio*, più o meno in potenza, più o meno fruibili nell'immediato, ma degni delle particolari attenzioni che da sempre riserviamo a questa fase del ciclo vitale. La dimensione storica e culturale dell'accesso va perciò considerata parte integrante del

[5] Federico Valacchi, «Quiddam divinum. Riflessioni sul metodo storico», *Archivi* XV, fasc. 1 (2020): 69–87.

lavoro di progettazione, evitando di correre il rischio di affidare il recupero delle informazioni a impalpabili sistemi di information retrieval che, per quanto potranno imparare, rischiano di restare troppo generici ed evasivi.
Se ammettiamo che i progressi del metodo possano riuscire a metabolizzare queste trasformazioni, negli archivi digitali nativi il problema del recupero degli atomi informativi tecnicamente non si pone perché in presenza di documenti digitali sarà sempre possibile operare una ricerca full text. A patto naturalmente che l'archivio sia costruito in modo tale da garantire sempre la contestualizzazione del dato e che sia appunto un archivio, non un grande contenitore di bit in ordine sparso. In questi archivi l'inventario, sempre ammesso che lo vogliamo chiamare ancora così, non è più *dell'archivio* ma *nell'archivio*, ne fa parte integrante come peculiare funzionalità di ricerca, con tutto ciò che ne consegue anche a livello di tassonomia e di definizioni degli strumenti nel loro insieme.
Diverso il caso delle acquisizioni di documenti o nuclei di documenti provenienti dall'enorme eredità analogica, cioè della cosiddetta digitalizzazione delle fonti primarie.
La prima questione da affrontare al riguardo è quella, decisiva, della contestualizzazione/ ricontestualizzazione, cioè del rapporto tra l'evidenza digitalizzata e l'integrità del fondo originario a cui si attinge. In presenza di processi di digitalizzazione selettiva l'ansia di costruire grandi serbatoi di "cose" digitali, dove la quantità vince sui sistemi di relazioni, può risolversi in rigenerazioni informative, se non in vere e proprie degenerazioni, le cui conseguenze possono essere piuttosto serie.
La dematerializzazione non si può negare, è semplicemente un dato di fatto. Si può però tentare di interpretarla, se non di governarla. Nel pieno di anni ipermnemonici, l'archivistica può dare il suo contributo a una più generale riflessione critica su quello che facciamo con le nostre tecnologie. Non si deve inventare nulla, basta seguire secolari processi virtuosi. L'ordine e l'inventario continuano ad essere i veri garanti di una coscienza critica dell'archivio, indispensabile soprattutto negli sviluppi non sempre coerenti della digitalizzazione. Nelle politiche dematerializzanti, quindi, prima dovrebbero arrivare gli inventari e poi gli *oggetti* che essi descrivono o introducono. Una digitalizzazione *object oriented*, senza adeguata descrizione e senza ordinamento preventivo, è un'anatra zoppa.
A prescindere da ogni altra considerazione, e dalla sua intrinseca qualità, nessuno degli strumenti attuali sembra però avere ad oggi la forza di superare i limiti non scritti della mediazione archivistica. Sono molto ma non tutto, si può fare di più.
Bisogna allora che il meccanismo faccia uno scatto, e si entri nel merito del recupero del dato all'interno del singolo oggetto digitale. Dobbiamo cioè sforzarci di spostare l'attenzione dai metadati ai dati, rilanciando, almeno in prima battuta, alcune strategie proprie del metodo per materia e basate in sostanza su particolari marcature del testo. L'indicizzazione e la soggettazione supportano già l'efficacia della gestione documentale e sostengono le attività di classificazione nella fase corrente. Queste tecniche possono dare il loro contributo anche nei fondi storici, soprattutto quando le si usi con la dovuta prudenza[6].
La usuale descrizione archivistica può essere incrementata da ulteriori metadatazioni, magari in forma di tag. Un tag, in questo senso, è un metadato che avvicina al contenuto, per quanto sia anch'esso una forma di interpretazione di chi lo genera.
Ci si può chiedere poi se esistano altre forme possibili di riconoscimento del testo e se si possa quindi spostare l'azione euristica dai metadati ai dati. Il problema di base è quello, noto da molto tempo, della difficoltà che una macchina incontra nel riconoscere nei segni dei significati lungo il processo di *handwritten text recognition.* Nello specifico, le tecnologie HTR nella loro costante evoluzione sembrano promettere risultati di sicuro interesse[7].

[6] Roberto Guarasci e Mauro Guerrini, *Cos'è l'indicizzazione* (Milano: Editrice Bibliografica, 2022).
[7] Denis Coquenet, Clement Chatelain, e Thierry Paquet, «Handwritten text recognition: from isolated text lines to whole documents», in *ORASIS 2021* (Saint Ferréol (France): Centre National de la Recherche Scientifique [CNRS], 2021), https://hal.science/hal-03339648

3. CONCLUSIONI

Una delle suggestioni più forti che i processi di dematerializzazione suscitano nel dominio degli archivi, e degli archivi storici in particolare, prende forma nella speranza di riuscire a disporre di documenti nei quali sia possibile operare puntuali ricerche per parola.

Le esorbitanti quantità informative con cui ci si confronta e l'indomabile anarchia della parola scritta rimangono però ostacoli di tutto riguardo. Ad oggi gli automatismi di ricerca passano ancora da un assiduo lavorio di trascrizione "manuale", finalizzato ad accrescere l'esperienza cognitiva della macchina per allenarla al riconoscimento dei segni. Trascrizione e verifica dei risultati dell'apprendimento sono gli strumenti di un lavoro tanto più moderno quanto antico e multidisciplinare.

Siamo a tutti gli effetti nel quadro di una descrizione archivistica aumentata, che non si ferma all'identificazione dell'oggetto, ma cerca di coglierne anche il contenuto, passando appunto da sistema di metadati a uno di dati contestualizzati. La descrizione identifica l'oggetto e la trascrizione lo svela, in un crescendo che sappia offrire al software materiale di confronto in grado di "allenarlo" e di potenziarne le performances cognitive specifiche.

Una descrizione che punti ai contenuti oltreché ai contesti si arricchisce delle trascrizioni di porzioni selezionate del fondo, affidandosi a una logica incrementale, sorretta dal learning machine e dalla sua possibile crescita cognitiva specifica.

Si pone certamente il problema della selezione e dell'alterazione dei vincoli costitutivi, con tutti i rischi che ne conseguono. In linea teorica la selezione contraddice il mito dell'avalutatività, ma lo stesso metodo che sembra porre dei limiti può risolvere la contraddizione. Se l'ordine conferito al fondo e il suo inventario ci tutelano, e se le finalità dell'azione sono esplicitate, niente proibisce infatti di pensare a approfondimenti "tematici" su porzioni del fondo. Si potrebbe cioè immaginare una postproduzione degli strumenti, arricchita da opzioni di ricerca capaci di spingersi in profondità non tanto nelle relazioni, ma nelle parole di cui ogni fondo alla fine è costituito. La tematizzazione non è reato, se è sostenuta da un'effettiva contestualizzazione e muove magari dall'analisi dei bisogni prioritari della ricerca che è possibile stabilire a partire dai comportamenti degli utenti. Passare da una generica iconografia digitale, fatta di immagini inerti, a una restituzione dinamica dei contenuti dovrebbe anzi essere uno degli obiettivi prioritari di una digitalizzazione virtuosa dei documenti di archivio.

Se volessimo recuperare il linguaggio degli standard, si tratta di passare dal potere incontrastato della multivellarità relazionale di ISAD(G) alla rete multidimensionale di significati e di rinvii logici e semantici di RiC[8]. Non è del resto un caso che proprio RiC, standard di seconda generazione, dichiari esplicitamente la sua dipendenza anche dalla tecnologia di cui disponiamo, cioè da una tecnologia sempre più duttile, potente e capace di sciogliere nodi atavici che arrivano direttamente dalla bidimensionalità descrittiva analogica.

Sono naturalmente ipotesi e soluzioni da verificare con tutte le precauzioni del caso ma che promettono sviluppi interessanti, che vanno anche oltre il recupero del dato secco.

Questo processo di crescita incrementale della conoscenza potrebbe in prospettiva fare affidamento anche sul dato quantitativo garantito dalle ricerche che si sviluppano dalle ricerche. L'approssimazione informativa si può infatti combattere non solo grazie all'intelligenza artificiale ma anche a quella "collettiva", tutta umana ed esperienziale.

In definitiva, quindi, prima ancora di immaginare le soluzioni tecnologiche, occorre abbracciare le logiche di una descrizione integrata, finalizzata alla costruzione dei sistemi interculturali che si affacciano anche dalle pagine di RiC.

[8] Giorgia Di Marcantonio, «Resource Description and Access e il modello concettuale Records in Contexts. A Conceptual Model for Archival Description: oggetti comparabili?», *JLIS.it* 9, fasc. 1 (2018): 128–35, https://doi.org/10.4403/jlis.it-12412; Pierluigi Feliciati, «Archives in a Graph. The Records in Contexts Ontology within the framework of standards and practices of Archival Description», *JLIS.it* 12, fasc. 1 (2021): 92–101, https://doi.org/10.4403/jlis.it-12675

Si tratta di capire se e in che modo l'esperienza archivistica riesca a dialogare con le altre discipline dell'informazione e sostenere l'apprendimento delle macchine, per aprire i documenti archivistici alle potenzialità di una ricerca puntuale e quanto possibile indipendente dalle rigidità gerarchiche.
In palio c'è la possibilità di svincolare i fondi archivistici dalla figura archetipica del soggetto produttore, per farli confluire dentro a quadri informativi più ampi e articolati, dove il confronto con le digital humanities si arricchisce di ragioni e di speranze descrittive.[9] E dove l'archivistica può recuperare un ruolo importante in contesti del tutto mutati ma non per questo da trascurare.

BIBLIOGRAFIA

[1] Coquenet, Denis, Clement Chatelain, e Thierry Paquet. «Handwritten text recognition: from isolated text lines to whole documents». In ORASIS 2021. Saint Ferréol (France): Centre National de la Recherche Scientifique [CNRS], 2021. https://hal.science/hal-03339648.

[2] Craig, Barbara Lazenby. «Old Myths in New Clothes: Expectations of Archives Users». Archivaria 45, fasc. Jenuary (1998): 118–26.

[3] Daquino, Marilena, e Francesca Tomasi. «Digital Humanities e Library and Information Science. Attraverso le lenti dell'organizzazione della conoscenza». Bibliothecae.it 5, fasc. 1 (2016): 130–50. https://doi.org/10.6092/ISSN.2283-9364/6109.

[4] Di Marcantonio, Giorgia. «Resource Description and Access e il modello concettuale Records in Contexts. A Conceptual Model for Archival Description: oggetti comparabili?» JLIS.it 9, fasc. 1 (2018): 128–35. https://doi.org/10.4403/jlis.it-12412.

[5] Feliciati, Pierluigi. «Archives in a Graph. The Records in Contexts Ontology within the framework of standards and practices of Archival Description». JLIS.it 12, fasc. 1 (2021): 92–101. https://doi.org/10.4403/jlis.it-12675.

[6] Gardini, Stefano. «Economie circolari dell'archivio: la carte di utenti e studiosi come archivi derivati». Nuovi Annali della Scuola speciale per archivisti e bibliotecari XXXV (2021): 237–77.

[7] Guarasci, Roberto, e Mauro Guerrini. Cos'è l'indicizzazione. Milano: Editrice Bibliografica, 2022.

[8] Romiti, Antonio. «I mezzi di corredo archivistici e i problemi dell'accesso». Archivi per la storia III, fasc. 2 (1990): 217–46.

[9] Tomasi, Francesca, e Marilena Daquino. «Modellare ontologicamente il dominio archivistico in una prospettiva di integrazione disciplinare». JLIS.it 6, fasc. 3 (2015): 13–38.

[10] Valacchi, Federico. «Quiddam divinum. Riflessioni sul metodo storico». Archivi XV, fasc. 1 (2020): 69–87.

[9] Francesca Tomasi e Marilena Daquino, «Modellare ontologicamente il dominio archivistico in una prospettiva di integrazione disciplinare», *JLIS.it* 6, fasc. 3 (2015): 13–38.