

On University Researchers Recruitment

Editorial board

Contact: editor@jlis.it

Received: 19 March 2026; Accepted: 20 March 2026; First Published: 15 May 2026

ABSTRACT

University recruitment constitutes a very delicate moment in the life of aspirants and in the very existence of the university. The main weakness lies in the request for a high number of publications, with articles published in class A journals. How is it possible that a research doctor between the ages of twenty-five and thirty can boast such a copious production? It follows that “Âgée” candidates are the winners. The late entry of researchers constitutes a problem of absolute relevance and emergency. It is no coincidence that the Italian university teaching staff is the oldest in Europe.

KEYWORDS

Recruitment; Researchers; University.

University recruitment, for LIS disciplines, is an extremely sensitive period in the lives of aspiring researchers and in the very existence of university. Nowadays, in Italy, the main weakness lies in the requirement for a high number of publications (12, and 15 for departments of excellence), with papers published in top-tier journals. How is it possible that a PhD graduate, in his mid-twenties or early thirties, can boast such a prolific and consistently high-profile output? Does it make sense to prioritize the quantitative criterion? The “publish or perish” philosophy leads to an overproduction of scientific literature not always of high quality, as well as to a hierarchical system where the publication venue is more important than the content of the article, since scientific journals are rigidly classified according to a value scale. A paper should be evaluated regardless of its publishing context, and the distinction between a department of excellence and a “normal” department has no sense. As a result, successful candidates are long-term candidates, who have had time to publish books and papers, and many of them are exhausted after years of precarious employment or professional work in archives and libraries. The late entry of researchers is a significant and urgent problem. It’s no coincidence that the Italian university teaching staff is the oldest in Europe. Recruitment methods vary across Europe, but a research project is evaluated both for the admission to a doctoral program (as in Italy) and for the admission to a research position. A candidate who, ideally, at 23-24 years of age wins a PhD, completes it three years later, still needs a couple of years to publish their thesis, to be obviously revised. This would require regular and generous funding for post-doc projects, where the doctorates could gain more experience and develop their own line of research, not necessarily emerging at the end of their PhD cycle. Therefore, the present university reform project, amid a substantial reduction in resources, risks penalizing postdocs and is a cause for concern.

This could be a better path to ensure a long term of activity for the PhD winners, thus gaining experience and developing the academic mindset necessary for effective advocacy within their own university and the scientific community. The research project solution, with a limited number of publications (certainly not 15), does not prevent indiscriminate participation, but at least allows young scholars to compete with hope of success. These are significant aspects that require analysis and further discussion within Italian academia, within AIDUSA and SISBB.

Sul reclutamento dei ricercatori universitari

Editorial board

Contact: editor@jlis.it

Received: 19 March 2026; Accepted: 20 March 2026; First Published: 15 May 2026

ABSTRACT

Il reclutamento universitario costituisce un momento delicatissimo della vita degli aspiranti e dell'esistenza stessa dell'università. Il vulnus principale sta nella richiesta di un numero elevato di pubblicazioni, con articoli editi in rivista di classe A. Com'è possibile che un dottore di ricerca a cavallo dei venticinque e trent'anni vanti una produzione così copiosa e tutta di alto profilo? Ne consegue che risultano vincitori candidati "âgée". L'immissione tardiva dei ricercatori costituisce un problema di assoluta rilevanza ed emergenza. Non casualmente il corpo docente universitario italiano è il più vecchio d'Europa.

PAROLE CHIAVE

Reclutamento; Ricercatori; Università.

Il reclutamento universitario per le discipline LIS costituisce un momento delicatissimo della vita degli aspiranti e dell'esistenza stessa dell'università. Il vulnus principale oggi in Italia sta nella richiesta di un numero elevato di pubblicazioni (12 e 15 per i dipartimenti d'eccellenza), con articoli editi in rivista di classe A. Com'è possibile che un dottore di ricerca a cavallo dei venticinque e trent'anni vanti una produzione così copiosa e tutta di alto profilo? Ha senso privilegiare il criterio quantitativo? La filosofia del *publish or perish* porta a una sovrapproduzione di letteratura scientifica non sempre di spessore, nonché a un sistema gerarchico per il quale vale più la sede in cui si pubblica il saggio rispetto a che cosa viene pubblicato, dal momento che le riviste scientifiche sono rigidamente classificate in una scala di valori. Un articolo va valutato indipendentemente dalla sede in cui viene pubblicato e non ha senso la differenza tra dipartimento di eccellenza e dipartimento "normale". Ne consegue che risultano vincitori candidati di "lungo corso" che hanno avuto il tempo di pubblicare libri e saggi, e diversi di loro arrivano esausti al ruolo dopo anni di precariato o di lavoro professionale presso archivi e biblioteche. L'immissione tardiva dei ricercatori costituisce un problema di assoluta rilevanza ed emergenza. Non casualmente il corpo docente universitario italiano è il più vecchio d'Europa.

In ambito europeo le modalità di reclutamento sono differenti, ma, in genere, viene valutato un progetto di ricerca sia per l'accesso al dottorato (come in Italia) sia per l'accesso alla posizione di ricercatore. Un candidato che, idealmente, a 23-24 anni, vince un dottorato e termina tre anni più tardi, ha ancora bisogno di un paio di anni di tempo per portare a maturità almeno la sua tesi, che va sempre rivista nella sua redazione. Ciò implica il finanziamento regolare e generoso di progetti post doc, in cui i neodottorati possano fare un po' di rodaggio ed elaborare una propria linea d'indagine che non necessariamente emerge alla fine del dottorato. Desta, pertanto, preoccupazione il complessivo progetto di riforma del sistema universitario che nel quadro di un sostanziale ridimensionamento delle risorse rischia di penalizzare proprio la fascia postdoc.

Questa potrebbe essere la strada da seguire per assicurare un lungo periodo di attività al vincitore, così da procurarsi esperienza e acquisire quella *forma mentis* accademica che è necessaria per compiere un'*advocacy* efficace all'interno del proprio ateneo e della comunità scientifica. La soluzione del progetto di ricerca, con un numero controllato di pubblicazioni (certamente non 15) non impedisce la partecipazione indiscriminata, ma almeno consente ai giovani studiosi di competere con qualche speranza di vittoria, aspetti non trascurabili e che necessitano di analisi e approfondimenti da discutere nelle nostre società scientifiche, AIDUSA e SISBB.

Many-to-many: the archival bond as a measure of potential relationships*

Federico Valacchi^(a)

a) University of Macerata, <https://orcid.org/0000-0003-2710-9316>

Contact: Federico Valacchi, federico.valacchi@unimc.it

Received: 15 January 2026; **Accepted:** 26 February 2026; **First Published:** 15 May 2026

ABSTRACT

The article examines the function of the archival bond and the different ways in which it manifests itself in processes of documentary sedimentation, and then proceeds to analyse the persistence of this foundational concept and its evolutions within the context of the dematerialisation of archives.

KEYWORDS

Archival bond; Archival relationships; Archival description; Archival sedimentation.

Molti a molti: il vincolo archivistico come unità di misura di relazioni potenziali

ABSTRACT

L'articolo valuta la funzione del vincolo archivistico e le diverse modalità in cui esso si manifesta nei processi di sedimentazione documentaria, per procedere poi all'analisi delle persistenze di questo concetto fondativo e delle sue evoluzioni nel quadro della dematerializzazione degli archivi.

PAROLE CHIAVE

Vincolo archivistico; Relazioni archivistiche; Descrizione archivistica; Sedimentazione archivistica.

* Questo articolo riprende, amplia e approfondisce temi emersi in occasione del convegno "Il vincolo archivistico. Relazioni, persistenze e trasformazioni nella sedimentazione documentaria" tenutosi a Cosenza il 30 e 31 ottobre 2025 e nella sintesi che se ne è fatta per gli atti in corso di pubblicazione.

1. Il senso del vincolo tra dogma e negazionismo occulto

Ogni archivio è il risultato di un sistema di pesi e contrappesi tenuti insieme da un'energia relazionale che poi i manuali definiscono vincolo archivistico.

Il vincolo, in quegli stessi manuali, è definito come un nesso determinante, spontaneo, necessario e involontario. Tuttavia, una simile definizione non è sufficiente a restituire la forza potenziale di un concetto che non si esaurisce nel semplice rilevamento di legami meccanici — talvolta quasi messianici — dai quali la *proprietà* dell'archivio sembrerebbe discendere automaticamente.

Questo nesso all'apparenza indissolubile, infatti, non è solo una discriminante teorica e logica che distingue l'archivio in senso proprio dalla raccolta, e neppure si riduce all'espressione elementare di una sequenzialità causale inevitabile. Il ruolo profondo del vincolo, al di là delle sue manifestazioni estrinseche più o meno evidenti, è piuttosto quello di determinare l'archivio come complicato sistema organico di dati qualificati e polifunzionali. Il bisogno che abbiamo avuto e abbiamo di sottenderlo alla definizione stessa di archivio in senso proprio, del resto, non è un ozioso esercizio di stile. Parlare di vincolo, infatti, significa prima di tutto preoccuparsi della qualità dei dati e della loro integrità. Se la vediamo così il vincolo non è una condizione, ma una garanzia e una promessa di qualità informativa.

Anche per queste ragioni, affrontare la sfaccettata questione del vincolo significa accettare il confronto con uno dei più delicati nodi metodologici della disciplina, per entrare con la dovuta accortezza nelle profondità sintattiche e logiche delle associazioni tra documenti. Il metodo in archivistica è peraltro un'inevitabile miscela tra presupposti teorici e osservazioni dei comportamenti documentari nei quali quel metodo si manifesta e trova, o non trova, conferma. L'esperienza, in ultima analisi, alimenta e sostiene una teoria che da sola non potrebbe reggersi, ma senza la quale la prassi faticherebbe a trovare un senso compiuto. Parlare di metodo significa quindi andare in direzione ostinata e contraria a una deriva teorica autoconsistente, per inseguire le evoluzioni della produzione e della sedimentazione, magari perimetrandone i confini, chiarendone le finalità e prevedendone gli esiti potenziali. Questo metodo non è presuntuosamente assertivo. È figlio di riscontri concreti e discuterne aiuta a individuare e a mettere a fuoco i comportamenti pratici e le loro conseguenze. Di conseguenza, il metodo allo stato puro non esiste, ma nelle sue molte possibili contaminazioni con la realtà rimane l'unica possibilità che abbiamo per parlare davvero di archivio. Il metodo è un punto di vista privilegiato. È un'aspirazione alla chiarezza, che ci serve soprattutto in una fase in cui si devono necessariamente fare i conti con i molti fraintendimenti innescati da un uso troppo spigliato del termine archivio.

Il vincolo archivistico, sotto molti punti di vista, è il simbolo e il baluardo di questo metodo dinamico e propositivo. È un'idea meravigliosa che scaturlisce da infinite speculazioni sul comportamento dei documenti, e non si è lontani dal vero se lo si considera come uno dei concetti fondativi dell'intera disciplina. Basta non esagerare e non farne un mito intangibile o addirittura un tabù. Basta, cioè, non aver paura di riconsiderarne il profilo al volgere di nuove stagioni e di nuove e inevitabili logiche della produzione documentaria.

Valutare il potenziale trasformismo del vincolo, ovviamente, non significa in nessun modo cercare di accantonarlo o di superarlo. Per inseguire il vincolo e le sue manifestazioni tanto "vincolanti", infatti, è sconsigliabile procedere per negazione. Non serve indulgere a certa ignorante iconoclastia o assecondare velleitarie rivoluzioni in un bicchiere d'acqua, combattute in nome di un panar-

chivismo sconsiderato. Non si deve neppure essere troppo tolleranti con chi risolve il problema della moltiplicazione delle fenomenologie documentarie dichiarando che tutto è archivio, tutto è archiviabile e tutto ciò che si ammassa “fa archivio”. Per quanto il polimorfismo debba essere valutato con attenzione e senza anacronistiche resistenze al cambiamento, infatti, c’è archivio e archivio, e alle distinzioni qualitative basate sui principi fondamentali è meglio non rinunciare. L’archivio, nella sua proprietà naturale e spontanea, è prima di tutto uno strumento del diritto, della certificazione e della garanzia e a queste prerogative fondamentali non si può abdicare. Chi parla di archivio trascurando le coordinate metodologiche che lo determinano crea solo i presupposti per alimentare l’enorme rumore di fondo che rimbomba sotto il cielo di questa parola ormai buona per tutte le occasioni.

L’archivio va soppesato nella sua specificità prima di trarre conclusioni frettolose. Bisogna ricondurlo alle sue leggi essenziali e ai suoi inossidabili principi fondativi. Esistono valori non negoziabili, la cui difesa non è una questione di orgoglio corporativistico. Si tratta piuttosto dello sforzo tenace di distinguere accumuli indistinti di informazioni da organizzazioni di dati qualificati in nome e per conto dei sistemi di relazioni che li inquadrano nella categoria immortale del contesto. Il punto di equilibrio, quindi, non sta nel reiterare un approccio mistico al concetto di vincolo e neppure in un’omissione ignorante del concetto stesso e delle sue conseguenze. Non si può essere favorevoli o contrari al vincolo, si tratta piuttosto di capire se, come e quando continui a manifestarsi.

2. Da subliminale ad object oriented: interpretazioni del vincolo archivistico

Come abbiamo già detto, gli archivi sono le prime vittime del loro polimorfismo innato e il vincolo che li sostanzia sembra destinato a seguirli sulla via delle trasformazioni, almeno nella sua accezione più consolidata. Per tentare di tracciare i cambiamenti, però, conviene sempre procedere in maniera comparativa e quindi non sarà inutile soffermarsi su alcune delle pagine che nel tempo la letteratura scientifica ha dedicato al nostro tema.

Partiamo col dire che il vincolo è parte integrante dell’archivio e in questo senso è un concetto senza tempo. Viene da molto lontano ed esisteva anche prima che lo si definisse teoricamente, magari in forme diverse da quelle in cui lo rappresentiamo oggi, come avremo modo di vedere. Da un punto di vista concettuale è però soprattutto durante il XX secolo che lo si è profilato per come lo conosciamo. Tutti i più grandi studiosi, in un modo o nell’altro, ci hanno fatto i conti e, anche se qui non ci interessa più di tanto ricostruire nel dettaglio questo lungo dibattito, ci sarà comunque utile richiamare alcune delle loro riflessioni, propedeutiche agli sviluppi del nostro ragionamento. Eugenio Casanova, intanto, non definisce il vincolo in quanto tale, ma lo fiuta in maniera quasi subliminale, colorata e maliziosa quando descrive le pratiche di archiviazione e i relativi strumenti di gestione, nella profonda consapevolezza del ruolo che le connessioni hanno nell’archivio. Scrive infatti il buon Eugenio: “il fascicolo (...) è la riunione ordinata per data o per numero degli atti ricevuti e spediti pel medesimo affare” (Casanova 1928, 142). E aggiunge, con taglio quasi pittorico, che quanto al fascicolo “insistiamo energicamente sul modo di costituirlo. Entro la coperta, il cui colore cambia secondo i titoli o categorie, gli atti devono essere collocati nell’ordine della loro compilazione e registrazione, lasciando quello col quale fu iniziata la pratica e quindi il fa-

scicolo, al centro, accavallandosi sopra esso i successivi; e facendo in modo che l'ultimo registrato rimanga a galla indichi subito all'esaminatore il punto ove sia giunta la trattazione della pratica e dunque un accavallamento alla rovescia regolato così dalla data dell'atto come dal nome assunto nella registrazione" (Casanova 1928, 147). Infine, con una nota non priva di una sapida malizia, Casanova precisa che "abbiamo espresso questa regola in forma condizionale, perché, purtroppo, salvo l'intitolazione del fascicolo, ben pochi uffici si curano di scrivere altro sulla coperta. Questa incuria permette la manomissione e toglie al fascicolo molto del suo valore giuridico. Immaginiamo soltanto quel che possa capitare in un fascicolo riservato del personale, quando l'interessato possa mettervi le mani sopra, né sia superiore a tutte le tentazioni" (Casanova 1928, 148). Già in Casanova, in definitiva, la sequenza fisica e logica della costruzione del fascicolo, inteso come insieme di documenti vincolati sia tra loro che con il produttore, rappresenta l'ordine dell'archivio, dell'amministrazione e della società.

Giorgio Cencetti, per parte sua, ribadisce e rinforza questa percezione quando lega i documenti nel vincolo della destinazione comune e nelle logiche, giuridiche prima che archivistiche, di un principio di provenienza trionfante. Scrive tra l'altro Cencetti: "I singoli componenti (...) non solo provengono dal medesimo individuo, aggregato familiare o ente (o dai suoi organi che torna lo stesso), ma poiché costituiscono niente altro che uno dei mezzi usati dall'ente o dall'individuo per raggiungere i propri scopi portano in loro stessi il vincolo della destinazione comune, sintetizzato nell'adempimento delle funzioni dell'ente o dell'individuo medesimo" (Cencetti 1937, 4).

L'universalità necessaria di Cencetti, in altre parole, è fatta di cose che hanno un senso solo se le si mettono in fila e il vincolo le tiene insieme, perché senza ordine relazionale non c'è archivio e senza archivio la società sbanda. La lettura di Cencetti irrigidisce l'archivio e limita la portata effettiva dei sistemi di relazioni e delle connessioni che collegano le componenti dell'archivio a livelli diversi. Quelle di Cencetti sono relazioni del tipo uno a molti, ma non si spingono fino al molti a molti che meglio definisce la complessità informativa di qualsiasi archivio. L'anelasticità di metodo e di merito del vincolo cencettiano viene colta prontamente da Claudio Pavone che, quando parla dell'archivio come organizzazione formale della memoria e dubita dell'efficacia meccanica del rispecchiamento, di fatto sposta il ragionamento su un terreno molto meno deterministico.

Questo limite latente viene definitivamente superato da Antonio Romiti. Il suo vincolo polimorfo e quadripartito, infatti, va oltre alla necessarietà aurorale che sottomette l'archivio al soggetto produttore, dentro a un rapporto di causa-effetto troppo rigido per essere vero. Romiti nei suoi raffinati laboratori isola infatti un vincolo articolato, che si potrebbe definire realistico e vitale, capace di abbracciare sostenere l'archivio in ogni sua manifestazione funzionale e strutturale. Come è ben noto abbiamo così il "*vincolo istituzionale esterno*", coincidente con l'insieme delle strutture istituzionali che appartengono a quei soggetti esterni che sono entrati in contatto con il soggetto principale, contribuendo alla complessa ed organica formazione dell'archivio"; il "*vincolo istituzionale interno*" rappresentato dalla struttura costitutiva del soggetto produttore; il "*vincolo archivistico esterno*" "che collega la documentazione conservata presso il produttore con quella che si trova presso tutti gli altri soggetti esterni che hanno avuto relazioni con il soggetto principale" e il "*vincolo archivistico interno*", cioè il nesso che collega, in maniera appunto logica e necessaria, tutta la documentazione costituente di un archivio di un soggetto (Romiti 2002, 49).

A ben guardare, Romiti ci parla di un vincolo ecumenico e commisurato all'ampiezza di un concetto di archivio che va ben oltre alla necessarietà meccanica delle sequenze documentarie. Al-

lude, cioè, a un vincolo che avvolge l'archivio inteso come espressione informativa complessa di contesti e contenuti. Se la leggiamo in controluce, poi, l'idea romitiana preconizza atteggiamenti e concetti che saranno propri degli standard e della loro forte vocazione alla strutturazione descrittiva. In più, e forse soprattutto, per quanto muovendosi ancora nella bidimensionalità cartacea, Romiti intravede l'anima dei sistemi informativi archivistici, con i loro altrettanto strutturati modelli relazionali interni ed esterni.

Le riflessioni di Romiti costituiscono ancora oggi un punto fermo e ad esse si è adeguata in sostanza tutta la letteratura successiva, che ha peraltro puntato molto sulla natura logica, necessaria e involontaria del vincolo, tanto che alla voce "vincolo archivistico" del glossario della Direzione Generale degli Archivi, fortemente influenzato dal pensiero di un'altra grande studiosa come Paola Carucci, si legge "nesso che collega in maniera logica e necessaria i documenti che compongono l'archivio di un ente".¹

Più recentemente sono emerse nuove riflessioni, in particolare quelle di Roberto Guarasci che, ragionando sull'evoluzione della natura e della struttura del documento e dell'archivio, ha posto il problema nelle sue conseguenze digitali. In particolare, almeno ai fini che ci proponiamo qui, sembra rilevante in Guarasci un approccio che di fatto ribalta la prospettiva, subordinando il vincolo all'oggetto digitale e alle sue manifestazioni. In questa lettura l'archivio come struttura formale e sostanziale non è più legittimato dalla sua coesione interna, possibilmente predefinita da un nesso statico e invariabile. Al contrario, il vincolo è cangiante. Di volta in volta reagisce alle modalità secondo le quali i documenti si organizzano per rispondere a un comando di visualizzazione sequenziale che diviene l'effettivo momento costitutivo del legame. Si chiede infatti Guarasci "se il nostro reiterato analizzare il vincolo archivistico non significhi che il vincolo nel digitale altro non è che una delle possibili visualizzazioni dei dati o dei documenti che compongono l'archivio, la cui connotazione non è più data dall'esistenza del vincolo ma bensì solo dalla produzione e proprietà del dato che esplica comunque degli effetti giuridici quando la visualizzazione assume alcune forme determinate".² Questo interrogativo, molto convincente sia nelle premesse che negli esiti, completa in qualche modo un percorso che va dal vincolo subliminale di Casanova a quello *object oriented* evocato dallo stesso Guarasci.

Come cercheremo di ribadire più avanti, l'idea che nel contesto digitale il vincolo sia un'opzione estemporanea legata a particolari accadimenti documentari è molto sensata. Introduce un punto di vista che ci consente di riflettere con maggiore puntualità sul presente e sul futuro del nostro beneamato vincolo, per coglierne la persistenza a prescindere dalla natura dell'archivio, ovvero per salutarne nuove manifestazioni. Per procedere in questa direzione, però, bisogna essere disponibili a ribaltare il rapporto tra il dogma da cui l'archivio in senso proprio dovrebbe discendere e le forme in cui l'archivio si manifesta invece in natura. Ci tocca, cioè, tornare a ragionare prima di tutto sull'archivio e sulla effettiva continuità nel tempo e nello spazio dell'involontarietà fondativa e delle forme che assume la necessità logica e relazionale.

¹ <http://2.42.228.207/archivi/index.php/abc-degli-archivi/glossario>.

² Questo passaggio è tratto da un contributo in corso di stampa negli atti del convegno "A proposito di archivi. Convegno in onore di Gianni Paoloni", tenutosi a Roma il 13 e 14 marzo 2025. Ringrazio il prof. Guarasci di avermelo anticipato e di aver acconsentito alla citazione.

3. Dal sogno dell'archivio alla volubilità elettrica

Il vincolo, per come ancora oggi in sostanza lo professiamo, brilla per necessità e involontarietà. Grazie a queste doti prodigiose benedice l'archivio che crea e da cui scaturisce. È il simbolo dell'archivio percepito con Derrida come “luogo dove l'ordine è dato”, e l'ordine, proprio per la sua sacralità giuridica e sociale, deve avere presupposti solidi e coesione interiore altrettanto robusta. In questo senso il vincolo è l'espressione manifesta di un bisogno per certi versi archetipico. Non è un tecnicismo, ma una forma di autenticazione che va ben oltre l'ordinata sedimentazione. Il vincolo tradizionale è il sigillo della fisicità di un archivio figlio di un'organizzazione che è gestionale ancora prima che genericamente archivistica e che si manifesta nel susseguirsi di documenti e fascicoli accuratamente legati tra loro e con il resto della sedimentazione.

Questo archivio accade dentro a reazioni a catena di natura certificatoria che risultano spontanee e naturali proprio in ragione della loro inevitabile necessità fattuale. Il vincolo, in questo senso, sta nell'ordine naturale delle cose o, meglio, dietro e dentro quest'ordine. Il vincolo presuppone, cioè, una sequenzialità necessaria scritta nelle cose e nel modo in cui le cose -ovvero le pratiche, ovvero i procedimenti- si sviluppano concretamente. In questo senso esso rappresenta il sostrato contestuale che alimenta una pratica in tutte le sue componenti ed è la garanzia di uno sviluppo armonico e tracciabile dei procedimenti. Il vincolo, detto in altre parole, almeno all'origine, è contesto probatorio più che astrazione relazionale, dal momento che la sua necessità qualifica l'archivio.

Questo vincolo sequenziale e, potremmo dire, giuridicamente rilevante, è un sistema raffinato per conferire un certo tipo di organizzazione a un archivio, partendo dal presupposto che l'archivio sia e comunque il risultato spontaneo dell'azione tracciabile e coerente di un soggetto produttore. Alle estreme conseguenze questo nesso è frutto di un pre-giudizio organizzativo di natura logica che si sostanzia nel titolare, cioè nel *somnium archivi*, cioè nella risposta precostituita al bisogno di organizzazione funzionale di un dato soggetto produttore. Il vincolo, in questo senso, scaturisce da una precisa realtà della sedimentazione.

La realtà, però, quando si parla di archivi, altro non è che una costante rincorsa alle continue trasformazioni della società e sotto questo punto di vista negli universi documentari niente rimane uguale a sé stesso troppo a lungo. Non si inventa nulla, perché se ricorriamo ancora alla *auctoritas* di Antonio Romiti, possiamo notare che “è opportuno precisare che la teorizzazione del concetto di vincolo non tiene sempre in considerazione le situazioni concrete e specifiche che si verificano necessariamente nella generale realtà archivistica, in considerazione delle differenti tipologie di archivio e della molteplicità delle vicende che arrecano con molta frequenza elementi di disturbo e di trasformazione strutturale generale” (Romiti 2002, 49-50).

Se ammettiamo questa volubilità innata e la naturale tendenza degli archivi a divincolarsi da modelli troppo rigidi, viene di nuovo da chiedersi se il vincolo sia un baluardo che resiste intangibile mentre il resto del mondo metodologico e funzionale va in pezzi, ovvero se possa essere oggetto di rivisitazioni anche piuttosto pesanti. È una questione non da poco, perché taglia in diagonale le fenomenologie archivistiche contemporanee e mette in gioco il forte polimorfismo che le contraddistingue, riqualificandone gli assetti relazionali.

Per valutare la questione nel modo giusto conviene prima di tutto fare un passo indietro e soffermarsi sulle possibili manifestazioni fisiche del vincolo e sulle loro conseguenze nella forma che

l'archivio assume. Come abbiamo visto fin da Casanova, un determinato concetto di archivio si basa sulla coesione documentale e gestionale di cui il vincolo è causa e conseguenza. Anche in questa lettura piuttosto dogmatica la sacra ineluttabilità del vincolo, come lo celebriamo talvolta nei fasti disciplinari, può però perdere qualche colpo, e non da oggi. Quando si prendono in considerazione modelli organizzativi diversi da quelli inchiodati al fascicolo qualcosa può cambiare. In una filza, e generalizzando in archivio di antico regime, il vincolo non è ad esempio un'astrazione concettuale o un sistema di relazioni logiche. È materia viva, e prende corpo in un pezzo di corda che lega fisicamente i documenti in sequenze cronologiche funzionali a un tipo di organizzazione che per prassi ed abitudine gestisce in quel modo la sedimentazione. In una filza il vincolo si può vedere e toccare, e perfino recidere con le mani.

In più, anche se circoscriviamo il ragionamento a un'idea di vincolo che è l'espressione del modello di gestione dell'archivio che si afferma agli inizi del XIX secolo, non mancano problemi e incongruenze tra il concetto astratto e la sua persistenza reale. Se infatti guardiamo agli archivi come entità in continua evoluzione, inserite in un complesso "tessuto" di relazioni e contesti, la natura di un documento e la sua posizione all'interno di un fondo archivistico non sono statiche, ma dinamiche, e dipendono da complessi legami contestuali (passati, presenti e futuri) e interconnessioni. Le vicissitudini della sedimentazione e della conservazione non ci parlano di strutture tautologicamente e invariabilmente "vincolate": scarti, scelte gestionali e accadimenti imprevedibili possono giocare la loro parte nell'alterare la presunta intangibilità del legame originario. Nell'auto rispecchiamento pavoniano il vincolo aurorale è stressato dagli avvenimenti conservativi e diventa qualcosa di diverso dalla sua ontologica intangibilità.

Tutto questo ci dimostra come nel tempo si siano già manifestate forme di vincolo diverse da quelle strettamente classificatorie e procedurali, perché qualsiasi modello gestionale è esso stesso il risultato di un'organizzazione e di un'interpretazione cui corrispondono distinti sistemi di relazioni. La questione non sembra irrilevante perché, come vedremo nel dettaglio, ci apre alla possibilità di valutare l'idea di vincolo proprio dentro a sistemi di relazioni fluide e potenzialmente mutevoli, abbattendo il monolite di una struttura data per sempre, e per sempre coesa nella sua sostanziale bidimensionalità: non separi l'uomo, o la donna, ciò che la classificazione ha unito.

Le cose possono infatti ingarbugliarsi quando il documento e l'archivio, come ci ha ricordato Guarasci, non sono più immanenti entità atemporali, date a prescindere, ma esistono solo in ragione di una visualizzazione che li evoca da spazi di memoria alieni e non necessariamente strutturati. In tutti questi casi il vincolo, nella migliore delle ipotesi, perde la sua univocità a vantaggio di una frammentazione a necessità variabile.

Questo è vero in particolare quando si prende in considerazione la natura più intima dei documenti digitali e delle loro associazioni relazionali, che dipendono più da sequenze di prompt reiteratamente palinogenetici che da una loro precisa posizione nel tempo e nello spazio. Vale subito la pena di notare, al riguardo, come la digitalità in cui siamo immersi, in ogni caso, non è una forma di gattopardesco trasformismo analogico che lascia le cose come stanno. Non è, cioè, un banale cambio di supporto, ma l'irruzione di logiche provenienti da mondi molto distanti da quelli analogici. Porta con sé radicali trasformazioni che investono in pieno la teoria, le pratiche e, soprattutto, le più consolidate abitudini archivistiche.

Per cominciare, l'idea granitica di soggetto produttore, che ha dominato la scena a lungo e a ragione, perde qualche colpo a vantaggio di modalità di azione collaborative e necessariamente più

elastiche, assecondate proprio dai processi di dematerializzazione. Se ne sono accorti anche gli standard, e non a caso RiC prende atto del fatto che “un singolo record o un singolo accumulo di record può essere creato congiuntamente da più individui o gruppi, simultaneamente e/o successivamente. Più persone o gruppi possono svolgere ruoli diversi in relazione a un record o a un insieme di record. Le persone creano e utilizzano i record, ma spesso ne sono anche l’oggetto. (...) L’emergere dell’editing collaborativo nell’ambiente digitale in rete ha portato ad altre complessità nel determinare l’origine e la proprietà dei documenti, poiché (...) molti documenti digitali hanno una complessa paternità multi parte e l’uso di servizi di archiviazione remota non completamente controllati dagli utenti solleva questioni di proprietà e custodia”.³

Se la vediamo così, in determinati scenari la moltiplicazione degli *agenti produttori* si sostituisce alla stabilità di un *creator* percepito come padre padrone dell’archivio. Nel contempo, la stessa sedimentazione perde la sua autoconsistente consapevolezza fisica, a tutto vantaggio di più sfumate e sfuggenti “archiviazioni remote”. L’archivio è un ospite un po’ disorientato, spesso fuori da un reale controllo dei produttori e collocato in spazi disco remoti, chissà dove e chissà perché (Guercio 2011).

Siamo nel cuore del problema della diversa fisicità degli archivi digitali, da cui deriva poi la difficoltà di stabilizzare a priori un archivio che consiste invece nel lampeggiare asincrono di flussi distribuiti in uno spazio di stoccaggio vago e per molti versi indifferente rispetto alla visualizzazione e alla consultazione. La consistenza materiale di questi archivi, che per altri versi è imponente e non priva di impatto, non è più inchiodata a spazi conservativi tangibili. Sono archivi che devono più all’apparire che all’essere, nei quali non c’è nessuna corda a dare contiguità fisica. In tutti questi casi l’archivio si disperde e si distribuisce in geografie diverse e distanti, non più inevitabilmente subordinate a quelle della produzione e dello stoccaggio. La manifestazione dell’archivio dipende dalle performance degli strumenti di interrogazione, ricerca, decodifica e restituzione. Di conseguenza, la necessità involontaria del vincolo, se resiste, è quanto meno a tempo determinato, e l’associazione logica ha logiche che possono modificarsi ad ogni interrogazione.

4. Le fluttuazioni del vincolo tra sistema archivio e classificazione secondaria conclusioni

Negli scenari che abbiamo evocato anche il ciclo vitale, dentro al quale sta scritto a chiare lettere il codice genetico dell’archivio analogico, subisce seri contraccolpi. Nel mondo digitale l’ipotesi di un tempo archivistico nettamente tripartito tende infatti da un lato a dilatarsi e dall’altro, come risulta sempre più chiaro, a collassare in un’implosione piuttosto fragorosa. Se si cerca di governare l’archivio digitale secondo la trimurti ancestrale, anche per la sopravvivenza di residui reperti analogici, ci si accorge ben presto che alle nostre tre età manca qualcosa. L’obsolescenza, come sappiamo bene, impone azioni conservative reiterate. Il ripetersi di una serie di procedure ineludibili si riflette sui tempi e sui modi della funzione archivistica. La conservazione non è più una responsabilità che si pone a posteriori: va progettata ed esercitata ciclicamente nel tempo (Pi-

³ International Council of Archives, *Records in Contexts foundations of archival description*, Version 1.0, November 2023 (<https://www.ica.org/resource/records-in-contexts-foundations-of-archival-description/>) p. 4. La traduzione è di chi scrive.

gliapoco 2016). La vita di un archivio digitale, quindi, non si risolve in un segmento temporale ben sporzionato, ma si sviluppa seguendo flussi periodici, rispetto ai quali il ciclo vitale tradizionale diventa al massimo un fattore di valutazione comparativa.

Se portiamo queste considerazioni alle loro estreme conseguenze, e se riflettiamo attentamente sulla natura fisica degli archivi digitali, insieme al vincolo e al ciclo vitale possiamo mettere in discussione la stessa percezione a compartimenti stagni del divenire conservativo. Si può, cioè, dubitare dell'utilità della creazione di un ambiente di deposito e di un archivio storico distinti fisicamente da quello corrente, riconsiderando l'effettiva sostenibilità funzionale di un periodico trasferimento di volumi documentari da un ambiente all'altro. Queste operazioni in prospettiva sembrano infatti ridondanti e piuttosto macchinose anche in termini banalmente applicativi. Pongono non pochi problemi di trasmissione, versamento, mediazione e fruizione e presuppongono una coerenza tecnologica e una puntualità archivistica di insieme che non di rado latitano presso i soggetti produttori e in qualche caso anche presso i conservatori, per quanto accreditati essi possano definirsi.

All'orizzonte si annuncia allora un ciclo vitale pesantemente rivisitato. Per meglio dire, si fa strada l'idea di superare un ciclo vitale "materiale", scandito dal tempo dei calendari o dal progressivo manifestarsi di finalità predeterminate, a vantaggio di partizioni logiche di un unico sistema, individuate dai diritti di accesso esercitati dalle diverse categorie di utenti. Il sistema di gestione documentale, insomma, si avvia ad essere anche l'ambiente di deposito e di conservazione, nella logica di un corpo documentario unico, alle cui specifiche componenti gli utenti potranno accedere in ragione dei diritti loro riconosciuti e delle rispettive finalità d'uso. Il ciclo vitale si costruisce in questo modo individuando a monte le categorie di utenti e i loro privilegi di accesso e ricomponendo in sostanza il big bang scaturito dalla rottura tra produzione, uso e conservazione da cui è nata l'archivistica come disciplina scientifica.

Tutto questo, sia detto per inciso, è ipotizzabile naturalmente solo nel caso di una riforma strutturale del modello conservativo che metta in equilibrio le diverse esigenze e dia le opportune garanzie in termini di attribuzione delle responsabilità e dell'esercizio delle azioni di tutela. Senza politiche archivistiche adeguate la deriva innescata dalla dematerializzazione sarà infatti incontrollabile e la nostra società sarà forse la prima nella storia a perdere quasi del tutto la sua memoria, e con essa anche il semplice piacere di ricordarsi.

Visto così, comunque, l'archivio diventa l'insieme delle cose, delle persone e delle azioni che scatenano processi di informazione e memoria tracciabili e conservabili nel tempo. Si avvia ad essere una rappresentazione estemporanea di ciò che le macchine, intese in senso ampio, collocano in una porzione di spazio e di tempo a sostegno delle esigenze degli umani e che determinati comandi richiamano in forma intellegibile ma non ultimativa. La provenienza non è più definita dal rapporto biunivoco e chiuso in sé stesso tra produzione e sedimentazione, ma va ricondotta concettualmente a un sistema complesso, fatto di impulsi e di azioni isolate e non solo di pacifiche istituzioni che sviluppano la propria attività in maniera lineare e gerarchicamente strutturata. Nelle pieghe dell'interoperabilità l'archivio digitale è una successione di eventi che scatenano processi documentari fluidi, elettrici.

Un archivio, e soprattutto un archivio digitale, non è per sempre, e non è sempre uguale a sé stesso. Sono i fatti a determinarlo, le azioni e i bisogni informativi di chi lo guarda a dargli evidenza (Rovella 2022). In questa lettura l'idea secca di archivio si diluisce e si complica nel concetto di

sistema archivio che oltre ai dati include le persone (intese anche come unità organizzative), gli strumenti e le procedure usate per dar luogo ai dati e utilizzarli, nonché i metadati e le diverse attività di autenticazione, gestione e conservazione che non sono esattamente l'archivio ma lo garantiscono (Valacchi 2006). Se lo si volesse definire in questa ottica l'archivio è il complesso dei dati, dei documenti, dei metadati, degli strumenti, delle procedure e delle persone generato, utilizzato e conservato da uno o più soggetti nell'esercizio delle proprie attività, anche in regime di interoperabilità, indipendentemente dalle tecnologie usate.

Definire così l'archivio ha però delle conseguenze non secondarie. Il sistema archivio, infatti, è una realtà dinamica e in costante ridefinizione. In questo ambiente le associazioni tra i documenti e i relativi vincoli si generano e si rigenerano con un certo tasso di variabilità. Di conseguenza, le rappresentazioni organizzative possono essere diverse da quelle precostituite del titolare che determina la classificazione archivistica standard, per quanto essa mantenga ancora un senso e anche il legislatore ne riconosca il ruolo. Si potrebbe parlare di una classificazione "secondaria" che scaturisce più da logiche di pertinenza che di provenienza in senso stretto. Questi flussi, in effetti, si creano intorno a un'esigenza informativa che può anche manifestarsi in forme diverse da quelle orgogliosamente sequenziali del fascicolo che rispecchia le funzioni del produttore. Il mondo dei documenti elettronici e delle viste documentali, coi suoi bisogni, tende a ragionare più per materia e pertinenza che per necessità e provenienza e torna a porre forte attenzione anche alla dimensione specifica, quasi diplomatistica, del documento archivistico (Guarasci e Folino 2013; Alfier 2020; 2025).

In tutti questi casi, infatti, il documento è protagonista in quanto "casa dei dati" e le sue associazioni temporanee tendono a prevalere sulla univocità fisica del complesso dato per sempre. Se è il comando "estemporaneo" a cercare il documento e a generare la struttura e la sequenza che poi vengono visualizzate, lo stesso concetto di vincolo assume un valore relativo e iterativo, alle soglie della sua negazione in quanto necessario e, soprattutto, involontario.

Quando l'archivio dipende dalla tendenziale soggettività della consultazione, e non viceversa, cambiano le carte in tavola, senza considerare che nella crescente interoperabilità dei produttori e dei sistemi il rigore silenzioso del titolare per funzioni tende a sbocconcellarsi, nel vano tentativo di inseguire logiche gerarchiche che l'essenza digitale, in sé combinatoria, tende a depotenziare.

A questo punto i giochi sono fatti e il vincolo non è più la lapidea fondazione dell'archivio, ma solo un sistema di relazioni effervescente e generativo.

Conclusioni

L'esigenza di confrontarsi con un archivio on demand, che si manifesta solamente quando se ne richiamano alcuni elementi costitutivi, spezza una lunga continuità d'uso e mette in crisi modelli di produzione, gestione e conservazione consolidati. La vittima più illustre di questo ribaltamento di prospettive è il concetto di vincolo, inteso come collante logico e fisico dell'archivio e del suo contesto probatorio immutabile.

Come abbiamo già visto, l'idea ancestrale di archivio è incatenata su un sistema di relazioni necessarie e involontarie di matrice logica e funzionale. In una visione simile il vincolo è percepito come uno degli elementi costitutivi del Creato, è qualcosa di innato e che esiste in sé prima dei tempi e

dell'archivio. In questo senso è indiscutibile e si tratta semmai di declinarlo nelle modalità che i singoli contesti documentari propongono.

Se assumiamo che l'archivio possa emergere da processi diversi da quelli più elementari della necessità universale la prospettiva cambia. La coerenza sequenziale dei sistemi gerarchici di relazioni vincolanti può essere sostituita da legami “deboli”, non previsti nella matrice precostituita.

Il vincolo polimorfo originario si alimenta di relazioni che in ultima analisi risultano prevedibili; quello “relativo”, invece, si manifesta in una più che potenziale proliferazione di legami, perdendo così la sua tranquillizzante funzionalità alla sedimentazione di archivi dotati di eccezionale potenza statica. Ciascuno di questi legami “irredentisti” introduce un carico informativo eccentrico, capace di modificare e riposizionare gli assetti complessivi del sistema archivistico. L'universalità, a questo punto, diventa anch'essa relativa, e oscilla ai ritmi con cui l'urgenza finalizzata dell'uso aggrega e distribuisce di volta in volta dati e documenti.

Se si adotta questo punto di vista, il vincolo altro non è che una possibilità o, meglio, l'inevitabile conseguenza di una determinata percezione di un archivio potenzialmente quantistico, che esiste solo quando lo si solleciti, un po' come il buon vecchio gatto dei manuali di fisica. La coerenza sequenziale dei sistemi gerarchici di relazioni vincolanti può essere sostituita da legami “deboli”, non previsti nella matrice precostituita. Il vincolo, dentro a tutta questa dinamicità, si trasforma in un set di metadati che sotto certi punti di vista ammicca ai poteri certificativi di una blockchain, orientato com'è a garantire tracciabilità, integrità e qualità informativa. In questa prospettiva, l'universalità del vincolo smette di essere necessaria per divenire relativa, oscillando al ritmo con cui l'uso funzionale aggrega e distribuisce dati e documenti.

Detto tutto questo, al vincolo, alla sua storia e ai suoi valori, nonché alla sua funzione, non possiamo né vogliamo rinunciare, perché comunque si manifesti esso è il segnale del necessario rispetto dell'essenza probatoria dell'archivio e in questo senso mantiene un irrinunciabile valore comparativo. Crediamo ancora nel bisogno di archivio. Siamo certi che tra un archivio, per quanto sempre più fluido, e una collezione ci siano differenze sostanziali, anche quando certi archivi assomigliano sempre più a collezioni o, forse, a collazioni. Continuare a riflettere sul vincolo non è quindi una forma di accanimento teorico, perché prendere davvero atto dell'esistenza di espressioni documentarie fisiologicamente diverse da quelle su cui si è costruito un intero universo epistemologico ha delle conseguenze. È urgente riclassificare il vincolo per non perderne la funzione fondamentale, a prescindere dalle forme dell'archivio e dalle modalità secondo cui il vincolo stesso si manifesta. L'archivistica è più che mai a un bivio. Se vuole sopravvivere come disciplina pensante ed evitare di diventare un oggetto di antiquariato, deve guardare in faccia la realtà per quello che è: un mondo digitale ineludibile e governato da leggi diverse da quelle analogiche. La continuità disciplinare è importante, ma il necessario salto di qualità sta proprio nell'abbandonare definitivamente l'approccio analogico alla dematerializzazione, per esplorare territori decisamente accidentati, e senza nemmeno indulgere a troppa prudenza, perché è davvero questione di vita o di morte.

I vecchi ferri del mestiere e l'armamentario teorico che ha avuto così tanta rilevanza vanno adeguati, perché ormai hanno solo un valore comparativo e, per quanto solidi, non sono stati pensati per reggere il confronto con la digitalità imperante. Il trucco, non nuovo per l'archivistica, sta nel leggere il presente con gli occhi del passato per prepararsi al futuro, in modo da difendere un sistema di valori fondativi al quale sarebbe molto pericoloso rinunciare.

In ogni caso, e comunque vada, ora e sempre viva il vincolo!

Riferimenti bibliografici

- Alfier, Alessandro. 2020. *Il sistema di documentazione digitale*. Milano: Editrice Bibliografica.
- Alfier, Alessandro. 2025. *Il documento e la sua autenticità. Uno strumento antico per la società digitale*. Milano: Editrice Bibliografica.
- Casanova, Eugenio. 1928. *Archivistica*. Siena: Lazzeri.
- Cencetti, Giorgio 1937. "Sull'archivio come universitas rerum." *Archivi* IV (1): 3-9.
- Guarasci, Roberto, e Antonietta Folino. 2013. *Documenti digitali*. Milano: Iter.
- Guercio, Maria. 2011. "Custodia Archivistica, Ubiquità Digitale." *Archivi & Computer* 2: 92-103.
- Pigliapoco, Stefano. 2016. *Progetto Archivio Digitale. Metodologia Sistemi Professionalità*. Torre del Lago: Civita Editoriale.
- Romiti, Antonio. 2002. *Archivistica generale. Primi elementi*. Torre del Lago: Civita Editoriale.
- Rovella, Anna. 2022. "From Archives Management to Knowledge Extraction: Tools and Impact Evaluation." *Comma* 1.
- Valacchi, Federico. 2006. *La memoria integrata nell'era digitale. Continuità archivistica e innovazione tecnologica*. San Miniato: Titivillus.

Measuring to improve: proposals for data collection and analysis in Italian archives

Marco Lanzini^(a)

a) University of Bergamo, <https://orcid.org/0009-0001-8975-1374>

Contact: Marco Lanzini, marco.lanzini@unibg.it

Received: 23 December 2025; **Accepted:** 09 March 2026; **First Published:** 15 May 2026

ABSTRACT

The essay examines the role of measurement and data analysis in the evaluation of archival services, highlighting delays, resistance and critical issues in comparison with other fields such as libraries and museums. Drawing on international standards and guidelines, it presents the recovery and reworking of historical data relating to the reading room services of the State Archives of Florence and Venice, which led to the development of a prototype web application for calculating metrics and generating tables and charts. The paper discusses in detail the limitations identified in the legacy data and proposes several measures to improve future data collection, advocating the integration of data gathering and analysis functionalities within the Sala Studio platform, which is gradually intended to manage user services for all Italian State Archives.

KEYWORDS

Data; Measures; Metrics; Archives; Evaluation.

Misurare per migliorare: proposte per la raccolta e l'analisi dei dati negli archivi italiani

ABSTRACT

Il saggio esamina il ruolo della misurazione e dell'analisi dei dati nella valutazione dei servizi archivistici, mettendo in luce ritardi, resistenze e criticità rispetto ad altri ambiti, come biblioteche e musei. A partire da standard e linee guida internazionali, vengono presentate le attività di recupero e rielaborazione dei dati storici relativi ai servizi di sala studio degli Archivi di Stato di Firenze e Venezia, che hanno condotto allo sviluppo del prototipo di un'applicazione web per il calcolo di metriche e la restituzione di tabelle e grafici. Nel contributo, si discutono in dettaglio i limiti riscontrati nei dati pregressi e si propongono alcuni interventi per migliorare le rilevazioni future, auspicando l'integrazione di funzionalità di raccolta e analisi dei dati all'interno della piattaforma Sala Studio, destinata a gestire progressivamente i servizi agli utenti di tutti gli Archivi di Stato italiani.

PAROLE CHIAVE

Dati; Misure; Metriche; Archivi; Valutazione.

1. I dati nella valutazione dei servizi archivistici tra prospettive, resistenze e criticità

Nel corso degli ultimi anni gli archivisti hanno iniziato a prestare maggior attenzione alle rilevazioni statistiche relative al patrimonio conservato, al suo uso e alle caratteristiche dell'utenza, pur muovendosi ancora in ordine sparso all'interno di un settore in cui non mancano resistenze e voci discordanti. Una parte della comunità archivistica si è ormai persuasa che solo attraverso l'uso di un linguaggio comune sia possibile ottenere risultati confrontabili, al fine di analizzare le caratteristiche dei singoli istituti e favorire un dialogo informato tra conservatori su come il patrimonio archivistico possa generare valore, nelle diverse accezioni del termine, e su come quello stesso valore possa essere incrementato (Hawk e Griffin 2022). Va in questa direzione la pubblicazione dello standard ISO 24083:2021 *Information and documentation – International archives statistics*,¹ che vedrà un ulteriore sviluppo con il futuro rilascio dello standard ISO/CD 25244 *Archives performance indicators*, finalizzato a trasformare i dati raccolti in indicatori di performance della qualità dei servizi archivistici,² documenti pensati per divenire punto di riferimento nella definizione di linee guida nazionali in grado di colmare il ritardo accumulato dal comparto archivistico rispetto a quello museale o biblioteconomico.

Il fatto che la valutazione dei servizi e i processi di *accountability* si debbano basare sull'evidenza dei dati è da anni un punto fermo di un mondo affine agli archivi, quello delle biblioteche, per cui esistono molteplici standard e *benchmark* ormai consolidati. Forte di questa tradizione, negli ultimi anni la biblioteconomia si è spinta a un livello successivo, affiancando alle indagini quantitative studi di tipo qualitativo volti a valutare l'impatto che la biblioteca può produrre nella società in cui opera e a cogliere l'immaginario che essa suscita nella comunità di riferimento, ricavandone ulteriori conoscenze "utili all'analisi e al monitoraggio dei servizi, alla loro progettazione e in definitiva al loro miglioramento" (Faggiolani 2019, 25). Nel considerare questo salto di qualità, Giovanni Solimine ricorda la circospezione con cui i bibliotecari della sua generazione già negli anni Settanta si spinsero sul territorio di confine rappresentato dalla rilevazione del grado di soddisfazione dell'utenza, nel quale "entravano in ballo variabili che non era facile dominare" con gli strumenti a disposizione, a differenza di quanto si era registrato per i processi valutativi basati su dati quantitativi, che rappresentavano già allora un "terreno 'certo', in cui si discuteva di danaro speso, di volumi acquistati o prestati, di utenti iscritti e così via" (Solimine 2019, 12).

Per gli archivi, l'intero processo appare più complesso da realizzare proprio a causa dei ritardi incontrati nella fase preliminare di acquisizione dei dati propedeutici alle successive indagini qualitative. Il terreno sotto i piedi dell'archivista, in tal senso, è ancor oggi friabile. Basti pensare al fatto che una porzione del patrimonio archivistico di ogni istituto, pur essendo regolarmente coinvolta nei servizi forniti all'utenza, dalla consultazione in sala studio al rilascio di copie, rimane in qualche modo sfuggente. Una lacuna, quest'ultima, dovuta alla complessa struttura di molti fondi, alcuni dei quali peraltro passibili di incrementi periodici che ne possono modificare la forma, e alla loro consistenza quantitativa, caratteristiche che raramente permettono di dotare

¹ L'anteprima dello standard è disponibile alla pagina web <https://www.iso.org/obp/ui/en/#iso:std:iso:24083:ed-1:v1:en>.

² Lo standard è allo stadio CD (Committee Draft), ma già in questa fase si pone esplicitamente "as a follow-up to ISO 24083", <https://www.iso.org/standard/89553.html>.

ciascuna unità archivistica di metadati descrittivi puntuali e standardizzati. Non di rado, intere porzioni di fondi sono prive anche solo di elenchi sommari che consentano di identificare le unità di condizionamento della documentazione (scatole, buste, faldoni, mazzi, etc.). Questi limiti non fanno che rafforzare in molti conservatori dubbi e resistenze rispetto a processi valutativi basati sull'evidenza dei dati, che richiederebbero, inoltre, l'adozione da parte dell'archivio di “systems to measure its performance against its plans”, come “recording systems, logs and paper-trails as well as numerical performance data” (NCA. PSQG 2008, 10).

Di fronte a queste difficoltà, gli archivi hanno spesso rinunciato a raccogliere e analizzare sistematicamente dati sul proprio funzionamento, benché quest'attività sia divenuta per molti versi più significativa da quando la mediazione tra il patrimonio e i suoi fruitori non è più necessariamente esercitata in maniera diretta dagli archivisti. Senza considerare le riproduzioni digitali di documenti consultabili online, per il momento limitate a una piccola percentuale della documentazione, negli ultimi anni sono aumentate le risorse descrittive fruibili attraverso sistemi informativi, portali tematici o anche solo i siti web delle singole istituzioni, che offrono banche dati e inventari in vari formati, dalla semplice scansione degli strumenti dattiloscritti e manoscritti alla produzione di pdf ricercabili. A uscirne minata è stata la possibilità degli archivisti di saggiare con mano bisogni ed esigenze del proprio pubblico, conoscenza che anche in passato, va ammesso, era in ogni caso parziale e fortemente influenzata da preconcezioni personali. All'aumento e al perfezionamento, reale o presunto, delle risorse archivistiche fruibili online corrisponde, inoltre, una riduzione delle occasioni in cui gli utenti possono fornire, anche inconsapevolmente, opinioni utili al miglioramento dei servizi.

L'esigenza di conoscere gli utenti attraverso indagini quantitative e qualitative appare ancor più stringente pensando alle prospettive di intermediazione algoritmica prefigurate dall'applicazione dell'intelligenza artificiale, che ridurranno ancor di più le occasioni di contatto tra conservatori e fruitori. Ancor più che in passato, in questo nuovo contesto il momento di confronto dovrebbe essere anticipato alla fase di progettazione di servizi e strumenti, con la partecipazione dell'utenza attraverso *focus group*, interviste, questionari e altre strategie di coinvolgimento (Feliciati 2012; 2016). Al tema ha dedicato una particolare attenzione il progetto internazionale di ricerca InterPARES Trust AI, all'interno del quale si è costituito un gruppo di lavoro su Reference and Access, con lo sviluppo di studi volti ad “acquisire dati attendibili sul punto di vista degli utenti, così da poter supportare (...) la definizione di requisiti e di linee guida per lo sviluppo di strumenti di intelligenza artificiale affidabili, sostenibili e utili per migliorare le funzioni di supporto e accesso a documenti e archivi” (Feliciati 2023, 122; Feliciati e Di Marcantonio, 2023). La raccolta e lo studio di dati sull'uso delle collezioni, in tal senso, può assumere una duplice valenza: in una fase propeudeutica alla progettazione di nuovi strumenti, per l'individuazione delle *personas* rappresentative del pubblico degli archivi da coinvolgere nel lavoro, e a valle della loro entrata in funzione, allo scopo di monitorare costantemente il comportamento e la soddisfazione dell'utenza rispetto ai servizi introdotti.

Pur con i ritardi e le difficoltà cui si è accennato, negli ultimi decenni, ancor prima del processo avviato con la pubblicazione e lo sviluppo degli standard ISO, non sono mancate, anche in ambito archivistico, iniziative che andassero al di là del contesto locale. Soprattutto nel mondo anglosassone, sono stati elaborati strumenti e linee guida volti a definire i dati da rilevare (misurazioni) e gli indicatori per monitorare l'evoluzione nel tempo di una singola misurazione o delle relazioni

tra di esse (metriche), con l'obiettivo di costruire *benchmark* chiari e condivisi. Risale a metà anni Novanta la nascita, in seno all'allora National Council on Archives del Regno Unito, del *Public Services Quality Group for archives and local studies*: un programma basato su sondaggi periodici finalizzati a saggiare il grado di soddisfazione degli utenti e raccogliere proposte per raffinare i servizi o implementarne di nuovi, con la definizione di uno standard nazionale per l'accesso agli archivi (Pickford 2002).³ Più articolato e in fase di espansione si presenta l'*Annual UK Archive Sector Survey*, con la raccolta di dati riguardanti istituti archivistici di diverso genere.⁴

L'esperienza britannica influenzò l'avvio del progetto *Archival Metrics and User Evaluation for Governmental Archives*, con il quale alcune università nordamericane hanno cercato di promuovere in ambito archivistico la cultura della valutazione basata sulla raccolta e l'analisi di dati relativi alle modalità di interazione tra istituti, utenti, personale d'archivio e patrimonio (Duff et al. 2010).⁵ Rimanendo negli Stati Uniti, di particolare interesse, soprattutto in relazione ai comportamenti dell'utenza, appaiono le linee guida *Standardized statistical measures and metrics for public services in archival repositories and special collections libraries*, licenziate tra il 2017 e 2018 dall'Association of college and research libraries (ACRL) e dalla Society of american archivists (SAA).⁶ Il documento prevede 8 aree: dati demografici degli utenti (1); attività di *reference* (2); visite in sala di consultazione (3); uso delle collezioni (4); eventi (5); formazione (6); esposizioni/mostre (7); interazioni online (8). Per ciascuna area vengono fornite le misurazioni di base che gli istituti dovrebbero raccogliere e quelle avanzate, da utilizzare sulla base di specifiche necessità dei singoli enti, con indicazioni relative alla natura e alle modalità di raccolta dei dati, alcuni esempi e le possibili metriche da sviluppare. A prescindere dalle difficoltà e resistenze di molti istituti nell'applicare concretamente le linee guida statunitensi – a causa della mancanza di strumenti informatici adatti, di risorse economiche e umane e, più in generale, delle competenze necessarie per raccogliere ma soprattutto per analizzare i dati –, la loro promozione ha contribuito a diffondere tra i conservatori nordamericani una maggior consapevolezza, anche solo sul piano ideale, dell'importanza che le rilevazioni statistiche possono assumere nei processi decisionali (Hawk e Griffin 2022).

Si tratta di un concetto che in Italia stenta a farsi strada, salvo alcune eccezioni limitate a specifici contesti locali o a particolari tipologie di istituti.⁷ Non molto diverso appare il panorama degli Archivi di Stato, al quale recentemente sono stati dedicati alcuni studi che, pur concentrandosi su specifici casi, forniscono un quadro esaustivo delle iniziative avviate in seno all'amministrazione archivistica italiana (Lanzini 2022; Gardini 2020; Gardini e Giacomini 2019). A fronte di informazioni attendibili sul patrimonio conservato, sulle strutture e sul personale – raccolte e pubblicate periodicamente anche grazie alla collaborazione tra MIC e ISTAT⁸ –, le stesse statistiche evidenzia-

³ I risultati dei sondaggi periodici sull'utenza degli archivi, attualmente curati dall'Archives & Records Association UK and Ireland, sono consultabili alle pagine web <https://www.archives.org.uk/resources> e <https://www.archives.org.uk/resources-archive>.

⁴ Si veda la pagina dedicata sul sito web degli Archivi Nazionali, <https://www.nationalarchives.gov.uk/archives-sector/our-archives-sector-role/annual-uk-archive-sector-survey/>.

⁵ Gli esiti del progetto sono pubblicati alla pagina web <https://sites.google.com/a/umich.edu/archival-metrics/>.

⁶ <https://www.ala.org/sites/default/files/acrl/content/standards/statmeasures2018.pdf>.

⁷ Un caso di studio di particolare interesse è rappresentato dal Censimento partecipato degli archivi storici in Emilia-Romagna - CASter (Chiarelli 2024, 76-8).

⁸ I dati relativi agli Archivi di Stato sono pubblicati sulla pagina web dell'Ufficio statistica del Ministero della cultura, <https://www.ministerocultura.gov.it/it/uffici/ufficio-statistica>.

no più di una falla in relazione ai servizi erogati all'utenza (Lanzini 2022, 385). Più in generale, le rilevazioni dell'amministrazione archivistica italiana paiono rispondere più ad adempimenti amministrativi che a una reale volontà di raccogliere dati in una prospettiva di analisi e valutazione dei servizi erogati. Uno dei principali limiti, riscontrabile tanto nei report annuali quanto negli studi più approfonditi, è rappresentato dal fatto che le informazioni a disposizione sono state per lo più raccolte per la gestione quotidiana dei servizi, senza un'adeguata attenzione a criteri, requisiti e procedure che le avrebbero rese più idonee a successive analisi statistiche. Non sorprende, quindi, riscontrare diverse forme di errore non campionario: dalla scarsa trasparenza dei valori censiti a classificazioni non corrette, sino ad alte percentuali di mancate risposte per una o più voci.

2. Il recupero dei dati pregressi degli Archivi di Stato di Venezia e Firenze e lo sviluppo dell'applicazione MAI

Soffriva di questi e altri limiti il saggio *Un'eco lontana da cogliere e interpretare. Proposte per un'analisi statistica dell'utenza degli archivi attraverso il caso di studio dell'Archivio di Stato di Milano* del 2022, che si concludeva con l'auspicio di poter ampliare l'indagine a due istituti in possesso di informazioni apparentemente omogenee a quelle milanesi, gli Archivi di Stato di Firenze e Venezia, accomunati dall'utilizzo dello stesso software applicativo per la gestione dei servizi di sala studio, Archimatic, entrato in funzione a Firenze nel 2002 e in seguito adottato a Milano (2007) e da ultimo a Venezia (2012). L'occasione per riprendere le fila del discorso giunse nel corso del 2024, quando Archimatic fu sostituito da Sala Studio, piattaforma gestionale dei servizi di sala studio integrata con il Sistema Informativo Archivistico - SIA, strumenti sviluppati dall'Istituto Centrale per gli Archivi nell'ambito di una revisione complessiva dei sistemi informativi archivistici ministeriali.⁹

In occasione della migrazione nel SIA dei dati di Archimatic relativi a fondi, collocazioni topografiche e indisponibilità dei pezzi, di cui si tratterà in altra sede, gli istituti coinvolti palesarono l'esigenza di poter continuare a interrogare anche le restanti informazioni pregresse, sia per effettuare verifiche sullo storico delle richieste da parte dell'utenza sia per finalità statistiche. Grazie a un accordo di ricerca con l'Archivio di Stato di Firenze e quello di Venezia, le esportazioni dei database originari (*dump*) sono state caricate in istanze PostgreSQL dedicate. Il progetto ha portato allo sviluppo di un'applicazione in Python, con l'uso della libreria Streamlit, per offrire agli istituti un'interfaccia web intuitiva attraverso la quale poter interrogare i database, con la produzione di grafici e tabelle esportabili nei formati CSV e XLSX. A differenza dello studio del 2022, condotto su dati aggregati prodotti attraverso le funzionalità di report di Archimatic, la nuova ricerca si è dunque basata su un set di misurazioni più ampio, attraverso il quale è stato possibile sviluppare metriche molto più complesse.

[tps://statistica.cultura.gov.it/?page_id=518](https://statistica.cultura.gov.it/?page_id=518), e nella sezione Cultura, comunicazione e viaggi della piattaforma IstatData, https://esploradati.istat.it/databrowser/#/it/dw/categories/IT1,Z0830COM,1.0/DCIS_ARCHIVIST. Gli stessi dati sono alla base del progetto sostenuto dal MIC e dalla Fondazione Scuola dei beni e delle attività culturali Minicifre della cultura, <https://minicifre.cultura.gov.it/>.

⁹ Per una descrizione dei due sistemi si vedano le relative pagine sul sito web dell'ICAR (Istituto Centrale per gli Archivi), <https://icar.cultura.gov.it/sistemi-e-portali/archivi-nazionali/sala-studio> e <https://icar.cultura.gov.it/sistemi-e-portali/archivi-nazionali/sia>.

Dopo il rilascio di una prima versione dell'applicazione, nel 2025 lo studio sui dati fiorentini e veneziani è proseguito con un duplice obiettivo: individuare eventuali ulteriori criticità nella struttura e nelle modalità di raccolta dei dati, opportunamente anonimizzati con la cancellazione delle informazioni personali;¹⁰ verificarne la compatibilità con le già ricordate linee guida *Standardized statistical measures and metrics for public services in archival repositories and special collections libraries*. In particolare, sono state considerate quelle aree del documento che più da vicino riguardano l'interazione tra utenza e patrimonio – dati demografici degli utenti (1); visite in sala di consultazione (3); uso delle collezioni (4) –, divenute fonte di ispirazione per lo sviluppo di una nuova versione dell'applicazione, ribattezzata MAI (Metriche Archivistiche Italiane), arricchita con alcune *query* che, attraverso una serie di filtri, restituiscono un'ampia gamma di informazioni. Le misurazioni e le relative metriche proposte dalle linee guida sono state adattate, per non dire piegate, al contesto degli Archivi di Stato italiani, non solo guardando alle problematiche riscontrate nei dati pregressi, ma anche tenendo conto delle criticità che difficilmente potranno essere superate nell'immediato futuro. La speranza, in tal senso, è che MAI possa fornire spunti di riflessione per lo sviluppo di nuovi strumenti di analisi statistica integrati in Sala Studio, oggi utilizzato da 17 istituti, numero destinato a crescere nei prossimi mesi.¹¹

Dati demografici degli utenti

Le linee guida statunitensi prevedono, come misurazione minima, la distinzione tra utenti interni o associati e utenti esterni. La rilevazione non si riferisce alla differenziazione dell'utenza dallo staff dell'archivio, che non dovrebbe essere preso in considerazione, se non quando usufruisce dei servizi a titolo personale, ma deve basarsi sulla definizione di elementi che creano un legame tra l'archivio e l'utente. Per un archivio comunale, ad esempio, potrebbero essere considerati interni tutti i residenti nel territorio comunale, mentre per un archivio universitario, tra gli interni possono essere annoverati studenti, professori e ricercatori dell'ateneo.

A partire da questa distinzione, vengono proposte due misurazioni avanzate riguardanti l'affiliazione e la registrazione. Nel primo caso, la classificazione può riguardare l'effettiva affiliazione istituzionale a un determinato ente, ma anche una più generica posizione lavorativa, il livello educativo, la residenza. Per quanto riguarda la seconda misurazione, invece, le linee guida suggeriscono di distinguere tra gli utenti che hanno usufruito dei servizi senza registrazione, come nel caso di risorse liberamente consultabili online, e quanti hanno ottenuto un permesso per accedere al patrimonio, fornendo una serie di informazioni utili alla loro classificazione. Le relative metriche

¹⁰ Nello sviluppo di MAI sono stati previsti due profili di accesso. Gli utenti interni agli istituti coinvolti, che accedono all'applicativo esclusivamente sulla rete intranet del Ministero della Cultura, possono utilizzare sia le funzioni statistiche su dati aggregati sia le funzioni di controllo e verifica a supporto della gestione dei servizi di sala studio, accedendo anche ai dati personali ereditati da Archimatic. Per eventuali utenti esterni, è stata implementata una seconda versione dell'applicativo, che limita l'accesso ai soli dati statistici: l'applicazione, in questo caso, si serve di copie di dataset preventivamente anonimizzati, ottenuti tramite rimozione dal database delle colonne non necessarie (e dunque potenzialmente identificative), così che i risultati siano costruiti su informazioni non personali e restituiti solo in forma aggregata.

¹¹ Gli istituti che a marzo 2026 si servono del sistema sono: Ascoli Piceno, Belluno, Cremona, Enna, Firenze, Milano, Modena, Napoli, Novara, Perugia, Perugia-Sezione di Assisi, Perugia-Sezione di Foligno, Prato, Roma, Salerno, Taranto e Venezia, <https://salastudio-archivi.cultura.gov.it/>.

possono ovviamente riguardare la variazione delle ‘famiglie’ di utenti nel tempo, sia in termini assoluti, sia in percentuale.

Visite in sala di consultazione e Uso delle collezioni

La terza area delle linee guida suggerisce di contare tutte le occasioni in cui un pezzo viene consultato e restituito da un singolo utente. Nel caso di una pausa tra mattinata e pomeriggio della stessa giornata, ad esempio, si dovrebbero registrare due distinte ‘visite’. In relazione all’uso delle collezioni, la misurazione minima proposta dalle linee guida conteggia la fuoriuscita del materiale dai depositi per qualsiasi tipo di utilizzo, dalla fruizione in sala studio alla movimentazione per attività di inventariazione, digitalizzazione, restauro, per eventi o lezioni, sino al prestito per mostre, consentendo di valutare l’impatto operativo prodotto da ogni spostamento.

In relazione all’uso di queste due aree, vanno fatte alcune precisazioni importanti. In primo luogo, i dati a disposizione hanno consentito di analizzare solo le movimentazioni destinate agli utenti di sala studio, non essendo presenti informazioni puntuali circa altri tipi di ‘utilizzo’ delle collezioni. La seconda questione, ancor più problematica, riguarda la mancata rilevazione in Archimatic del numero effettivo di visite in sala di consultazione, dato raccolto di norma in registri cartacei o con strumenti informatici distinti dal gestionale. Il concetto di ‘visita’, dunque, è stato sostituito con quello di ‘richiesta’ di un pezzo archivistico, pur nella consapevolezza che i documenti conferiti a un utente possono essere consultati per pochi minuti o, al contrario, per più volte in giornate diverse.

3. Analisi dei dati a disposizione

Nel considerare i database originari di Archimatic, le caratteristiche demografiche degli utenti sono state ricavate dalla tabella *Iscrizione*, mentre per la terza e quarta area delle linee guida si è dovuto necessariamente procedere a un compromesso, servendosi, in particolare, delle tabelle *Richiesta* e *Nodo*.

La tabella “Iscrizione”

La tabella *Iscrizione* contiene sostanzialmente tutti i parametri per classificare l’utenza. Di seguito si riportano i campi presi in considerazione in questa fase del progetto e quelli giudicati potenzialmente utili in prospettiva di sviluppi futuri, ma per il momento non utilizzati, con una breve descrizione e le percentuali di rilevazione da parte dei due istituti.

Tabella 1. Dati della tabella *Iscrizione* dell'Archivio di Stato di Venezia e dell'Archivio di Stato di Firenze

Tabella "Iscrizione"			
Campo	Descrizione	% ASVe	% ASFi
ID	ID univoco dell'iscrizione.	100	100
ID utente	ID univoco dell'utente, il cui nome e cognome sono riportati nella tabella <i>Utente</i> , non presa in considerazione.	100	100
Data iscrizione	aaaa-mm-gg	100	100
Data scadenza iscrizione	Campo non utilizzato. La data di scadenza è fissata alla fine dell'anno solare di iscrizione (Venezia) o 12 mesi dopo la data di iscrizione (Firenze).	100	100
Cittadinanza	Valori basati sulla tabella <i>Paese</i> .	99,99	96,97
Data di nascita	aaaa-mm-gg	99,31	98
Città di residenza	Campo non utilizzato. A testo libero.	98,41	96,58
Nazione di residenza	Valori basati sulla tabella <i>Paese</i> .	99,85	86,64
Provincia di residenza	Campo valorizzato con la sigla della provincia, limitatamente al territorio italiano, compilato solo parzialmente. Si riporta la percentuale relativa alle sole iscrizioni di utenti residenti in Italia, con la percentuale complessiva tra parentesi.	82,42 (65,53)	81,06 (59,16)
Titolo di studio	Campo non utilizzato. Valori: scuola dell'obbligo; diploma di laurea; laurea (primo livello); scuole superiori.	5,41	96,09
Professione	Valori: altro; architetto - ingegnere; archivista; docente universitario; dottorando - specializzando; funzionario MIBACT [attualmente MIC]; funzionario - impiegato; geometra; giornalista; imprenditore - commerciante - artigiano; in attesa di occupazione - casalinga; insegnante; libero professionista; militare; operaio - agricoltore; pensionato; religioso - sacerdote; ricercatore libero professionista; studente.	73,92	98,65
Azienda	Campo non utilizzato. A testo libero, compilato solo parzialmente.	2,2	45,89
Motivo della richiesta	Valori: interesse culturale personale; ricerca per attività scientifica; ricerca per finalità amministrativa - professionale; tesi di laurea; tesi di dottorato; esame universitario.	78,94	96,99
Ambito cronologico	Campo non utilizzato. Valori: fino a XXII [secolo]; XIII-XIV; XIV-XV; XVI-XVIII; XIX-XX; lungo periodo.	85,23	95,52
Ambito tematico	Valori: altro; archivistica - storia degli archivi; demografia storica; genealogia - storia familiare - araldica; geografia storica - storia del territorio; paleografia - diplomatica; storia antica - archeologia; storia contemporanea; storia del diritto; storia del Rinascimento; storia della cultura - istituzioni educative e culturali; storia della letteratura - filologia - storia della lingua; storia della musica - spettacolo - teatro; storia della proprietà - ricerche catastali; storia della scienza - tecnica; storia dell'architettura - urbanistica - restauro architettonico; storia dell'arte - del restauro; storia economica; storia locale; storia medievale; storia moderna e del Risorgimento; storia politica; storia religiosa; storia sociale - del costume - delle donne; storia territoriale.	81,04	96,99
Titolo ricerca	Campo non utilizzato. A testo libero.	99,56	81,04

Particolarmente utile, a fini statistici, si è rivelata la scelta di assegnare a ogni individuo un identificativo univoco, con la possibilità di valutare le variazioni nel tempo dei dati dichiarati a ogni iscrizione. Rispetto alla durata di quest'ultima, tuttavia, si segnala la particolarità del caso fio-

rentino: mentre a Venezia era prevista una registrazione dell'utente legata all'anno solare, con la conseguente scadenza al 31 dicembre anche delle iscrizioni effettuate pochi giorni prima, in linea con quanto avviene nella maggior parte degli Archivi di Stato, a Firenze l'iscrizione durava esattamente 12 mesi, circostanza che ha richiesto alcuni accorgimenti nel calcolo di misurazioni e metriche riguardanti l'utilizzo delle collezioni.

In merito alle diverse voci, raggiungono percentuali elevate i dati su cittadinanza e data di nascita, entrambi basati su liste di valori controllati che hanno semplificato l'individuazione di raggruppamenti di particolare interesse (*cluster*). Si riscontrano comportamenti molto diversi tra i due archivi in relazione a una serie di misurazione 'avanzate', tutte basate su liste controllate: titolo di studio, professione, motivo della richiesta, ambito cronologico, ambito territoriale e ambito tematico. In alcuni casi le informazioni sono talmente lacunose da non consentire valutazioni significative, come avviene per il titolo di studio, che a Venezia scelsero di non registrare sistematicamente, tanto da figurare solo per poco più del 5% del campione.

Presenta invece qualche criticità la misurazione della residenza, divisa in tre campi: nazione; provincia; città. Il dato più completo è rappresentato dalla nazione, mentre risulta parziale quello riguardante la provincia, presente in circa 4/5 del campione italiano. Le informazioni sulla città, sebbene presenti in percentuali superiori, al momento non sono state utilizzate perché derivanti da un campo a testo libero soggetto a errori di varia natura e all'uso di criteri di compilazione eterogenei. In questo caso, dunque, sarebbe necessario prevedere una normalizzazione dei valori, processo che permetterebbe anche di colmare le lacune relative alla provincia. Sarà questo un aspetto su cui si potrà tornare in futuro, anche grazie alle opportunità fornite dai più recenti sviluppi dell'elaborazione del linguaggio naturale e dei modelli linguistici di grandi dimensioni, in grado di estrarre, normalizzare e collegare semanticamente informazioni testuali. Per il momento queste procedure sono state solo testate, ma non integrate ufficialmente nella versione di MAI messa a disposizione degli istituti.

Le tabelle “Nodo” e “Richiesta”

La tabella *Nodo* descrive le aggregazioni logiche che compongono i diversi complessi e le loro relazioni gerarchiche. Si presentano di seguito i campi presi in considerazione, senza riportare le percentuali di compilazione da parte dei due istituti, trattandosi di campi obbligatori.

Tabella 2. Dati della tabella *Nodo* dell'Archivio di Stato di Venezia e dell'Archivio di Stato di Firenze

Tabella “Nodo”	
Campo	Descrizione
ID	ID univoco del nodo.
Descrizione	Denominazione del nodo.
ID padre	ID del nodo padre. Rimane vuoto in caso di nodo di primo livello.
Segnatura parziale	A partire dal nodo di primo livello, fornisce le denominazioni di tutti i nodi della gerarchia, sino al nodo descritto nel record, divise da virgola.
ID dell'albero	ID dell'albero al quale appartiene il nodo.

Per quanto riguarda la tabella *Richiesta*, un dato potenzialmente interessante riguarda la tipologia. Archimatic consentiva di gestire sia le richieste dei pezzi fisici (lettera p), sia quelle delle riproduzioni in microfilm, nelle quali furono ricomprese anche le consultazioni delle riproduzioni digitali (lettera m). Questa distinzione fu di fatto utilizzata solo a Firenze, così come avvenne per la terza tipologia prevista, riguardante le richieste effettuate dal personale dell'istituto (lettera f). Si tratta di specifiche che per il momento non sono state prese in considerazione, facendo confluire tutte le richieste in un unico calcolo, ma che in futuro potranno portare ad analisi più raffinate sulle diverse modalità di consultazione del patrimonio.

Due informazioni sempre presenti sono quelle relative alla data di richiesta e all'ID del nodo del complesso al quale si riferisce la richiesta, che rimanda al relativo record presente nella tabella *Nodo*. Per quanto riguarda il numero del pezzo richiesto, o quello della bobina nel caso di microfilm, per Firenze sono presenti ben 17.875 record vuoti, pari al 2,44% del totale. Si tratta di pezzi relativi alle serie di alcuni fondi (*Diplomatico* e *Notarile*) per le quali esistevano modalità di richiesta particolari, gestite in MAI attraverso procedure specifiche.

Tabella 3. Dati della tabella *Richiesta* dell'Archivio di Stato di Venezia e dell'Archivio di Stato di Firenze

Tabella "Richiesta"			
Campo	Descrizione	% ASVe	% ASFi
ID	ID univoco dell'iscrizione.	100	100
Tipo richiesta	Campo non utilizzato. Tipo di richiesta indicata da una lettera: f (richiesta effettuata dagli operatori dell'istituto per conto di un utente); m (richiesta di materiale microfilmato/digitale); p (richiesta di un pezzo). Di lato si riportano le percentuali per ciascuna tipologia di richiesta, a dimostrazione del diverso utilizzo da parte dei due istituti.	100 f = 0,005 m = 0,23 p = 99,76	100 f = 2,03 m = 15,25 p = 82,72
ID utente	ID univoco dell'utente, il cui nome e cognome sono riportati nella tabella <i>Utente</i> , non presa in considerazione. Consente di collegare la richiesta ai dati dell'utente riportati nella tabella <i>Iscrizione</i> .	100	98,25
ID nodo	ID univoco del nodo del complesso archivistico consultato. Consente di collegare la richiesta ai dati del nodo riportati nella tabella <i>Nodo</i> .	100	100
Data richiesta	aaaa-mm-gg. Non è stata presa in considerazione l'ora della richiesta.	100	100
Pezzo	Segnatura del pezzo richiesto.	100	84,05
Bobina	Segnatura della bobina richiesta. Per Firenze il riferimento alla bobina è di norma alternativo a quello che riguarda il pezzo, anche se in molti casi sono presenti entrambi i valori; a Venezia il numero di bobina è sempre affiancato dal numero del pezzo.	0,23	15,07 Alternativo al pezzo = 13,51 In aggiunta al pezzo = 1,56
Inserto	Di norma associato alla segnatura del pezzo, di cui l'inserto rappresenta una partizione.	0	2,43

Sono necessarie alcune precisazioni in merito al concetto stesso di pezzo e inserto. In linea di massima, in Archimatic l'inserto, utilizzato solo a Firenze, serviva per indicare qualsiasi partizione interna al pezzo, inteso per lo più come unità di condizionamento. L'inserto, dunque, corrispondeva

al fascicolo o al singolo documento. La stessa funzione semantica, tuttavia, poteva essere assunta dal pezzo stesso, termine generico che anche all'interno del medesimo istituto veniva utilizzato di volta in volta con significati diversi, a seconda delle modalità di consegna del materiale. Mentre a Firenze, ad esempio, il *Diplomatico* viene consultato richiedendo le singole pergamene, a Venezia e Milano il conferimento dei fondi membranacei avviene a livello di unità di condizionamento, in ciascuna delle quali l'utente potrebbe rinvenire decine di pergamene di suo interesse. In mancanza di un lessico controllato condiviso, non è dunque possibile avventurarsi nella compilazione di 'classifiche' sugli istituti di 'maggior successo', anche ammesso che qualche conservatore abbia questa tentazione. Al momento, i dati presi in considerazione sono utili se confrontati sul lungo periodo all'interno di ciascun archivio.

In sintesi, limitatamente alle elaborazioni statistiche, MAI opera quindi su un insieme di entità che costituisce un sottoinsieme operativo dello schema concettuale di Archimatic. In particolare, il sistema utilizza le entità *Utente*, *Richiesta*, *Nodo* e *Iscrizione*, secondo lo schema di seguito riportato.

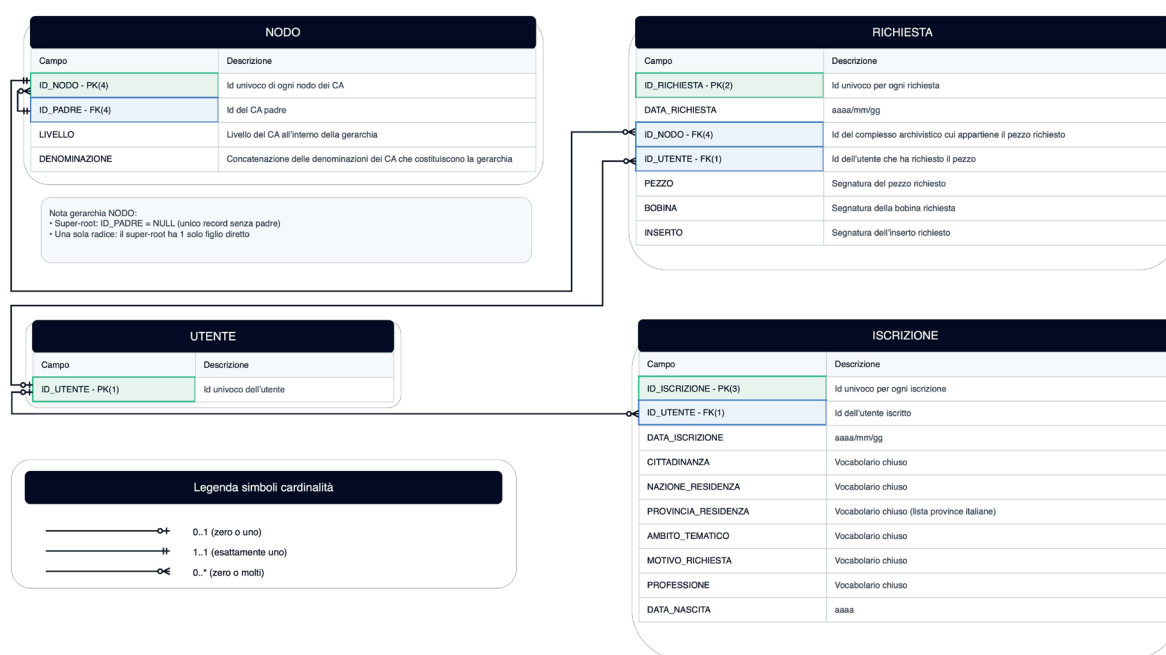


Figura 1. Schema Entità-Relazioni di MAI (nella versione limitata alle statistiche)

4. Metriche sviluppate e criticità riscontrate

Gli esempi riportati di seguito hanno finalità prevalentemente esemplificativa: servono a mostrare il tipo di interrogazioni e di letture rese possibili da MAI più che a proporre, in questa sede, un'interpretazione sistematica dei fenomeni.

Iscrizioni

Attraverso i dati illustrati sinora è stato possibile sviluppare all'interno di MAI diversi flussi di elaborazione (*pipeline*) che restituiscono il numero di iscrizioni per un determinato periodo, con i

seguenti filtri opzionali: provenienza (italiani; stranieri; stranieri residenti in Italia; italiani residenti all'estero; non dichiarato); cittadinanza; nazione di residenza; provincia di residenza; professione; fascia di età; motivo della richiesta; ambito tematico di ricerca. I risultati vengono mostrati sia attraverso una tabella, con la ripartizione per anno e mese delle iscrizioni complessive e di quelle di nuovi utenti, sia attraverso i relativi grafici (Fig. 2). L'applicazione, inoltre, produce altri grafici a torta e tabelle nei quali sono presentate le percentuali di iscrizioni sull'intero periodo e per anno, calcolate sulla base di parametri quasi identici ai filtri, ma non concatenabili tra loro: provenienza; cittadinanza; nazione di residenza; provincia di residenza; professione; fascia di età (0-17; 18-25; 26-40; 41-50; 51-60; 61-70; ≥ 71); motivo della richiesta; ambito tematico di ricerca (Fig. 3).

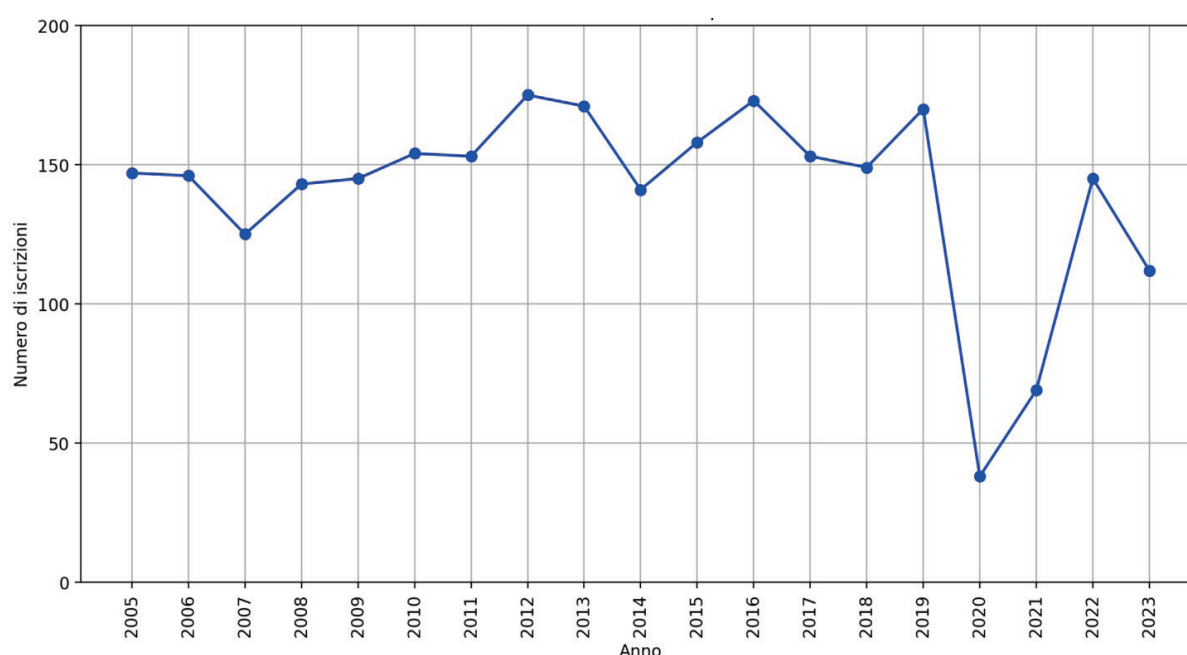


Figura 2. Grafico degli utenti iscritti all'Archivio di Stato di Firenze, con età tra i 18 e 25 anni, per il periodo 2005-2023

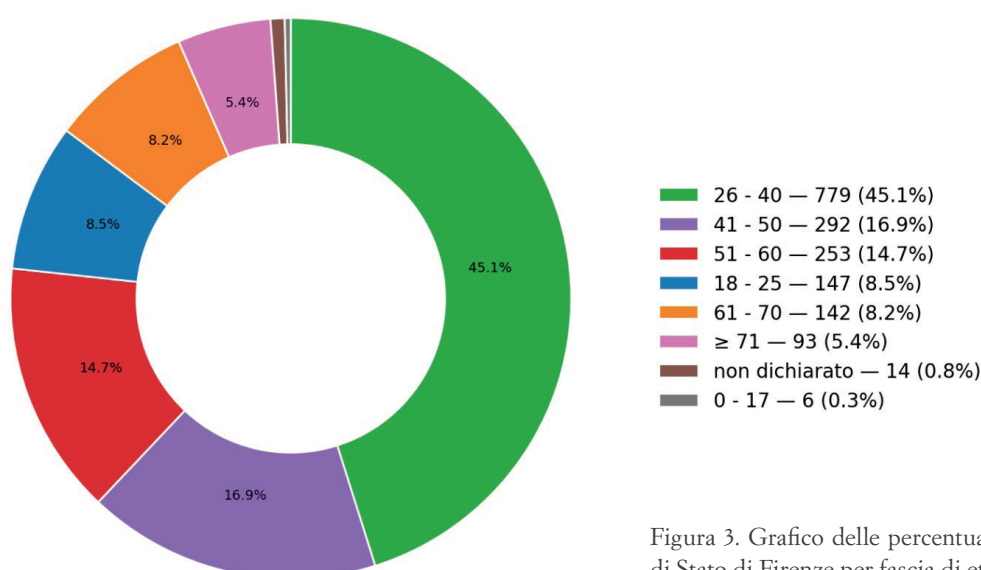


Figura 3. Grafico delle percentuali di iscritti all'Archivio di Stato di Firenze per fascia di età nell'anno 2005

La classificazione degli utenti può trasformarsi in uno strumento efficace per comprendere i punti di forza o debolezza di un istituto in termini di attrattività, ma da utilizzare con cautela. Un problema difficilmente sormontabile, ad esempio, deriva dal fatto che lo stesso utente potrebbe usufruire dell'Archivio per ragioni diverse da quelle dichiarate al momento dell'iscrizione, andando a inficiare le statistiche. Se alcune informazioni sono per loro natura stabili o comunque soggette a variazioni statisticamente irrilevanti nel corso di dodici mesi – si pensi alla cittadinanza –, altre possono mutare con una frequenza maggiore. Non è raro, soprattutto per utenti che frequentano un istituto assiduamente, iscriversi per avviare una ricerca relativa a un determinato tema e nei mesi a seguire effettuare, magari occasionalmente, ricerche riguardanti altri argomenti, che in questo modo non vengono in alcun modo considerati nella rilevazione.

Richieste

Il secondo gruppo di metriche gestite da MAI si riferisce alle richieste dei pezzi durante un determinato periodo. In questo caso, la misurazione può riguardare tutti i fondi dell'archivio, un singolo complesso o una sua parte. Nel calcolo possono essere applicati gli stessi filtri visti per le iscrizioni, con la produzione di tabelle e grafici per periodo, anno e mese (Fig. 4). Lo stesso discorso vale per i parametri attraverso cui vengono ricavate le percentuali delle richieste, con la restituzione dei risultati per periodo e singole annate. L'attribuzione dei valori utilizzati per filtrare le richieste segue regole che è opportuno illustrare nel dettaglio per ragioni di trasparenza. Di norma le caratteristiche dell'utente che richiede un pezzo sono ereditate dall'iscrizione relativa all'anno in cui viene effettuata la richiesta o, in caso di assenza del relativo record, da quella dell'anno precedente, comportamento che ha consentito di gestire il caso fiorentino illustrato in precedenza, dove esistono numerosi utenti che si iscrivono in un anno e continuano ad accedere all'istituto anche nell'annata seguente, senza la necessità di rinnovare l'iscrizione a partire dal 1° gennaio.

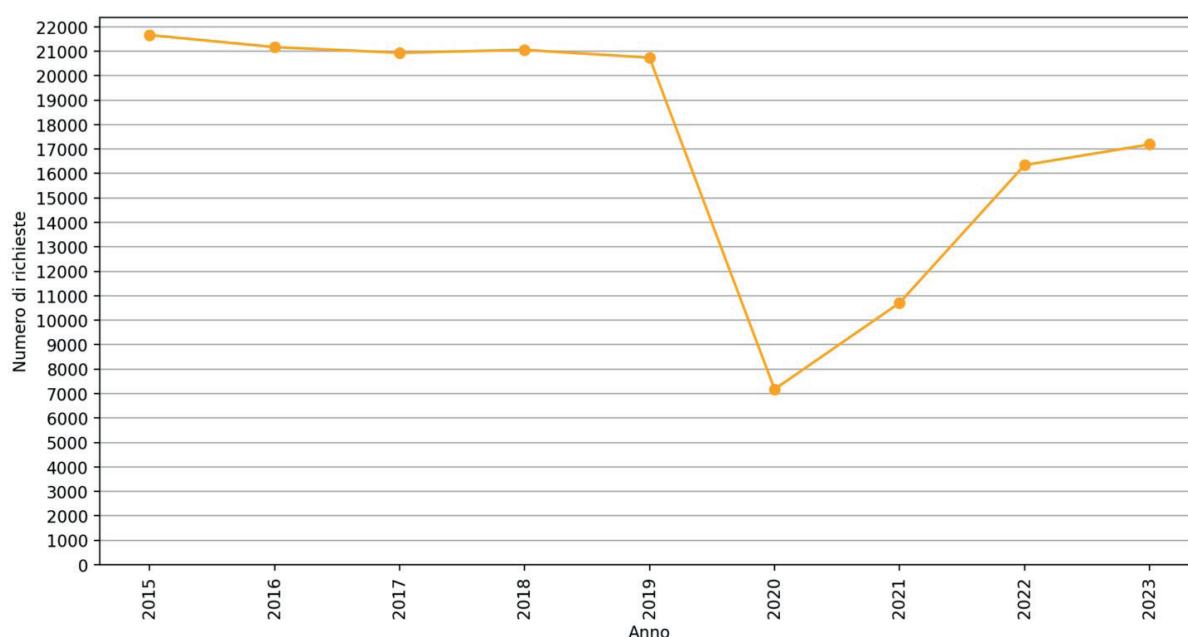


Figura 4. Grafico dei pezzi richiesti dagli utenti dell'Archivio di Stato di Venezia per il periodo 2015-2023

Nei casi di ulteriore esito negativo, il sistema procede a ritroso per cercare un'eventuale iscrizione precedente e, se non la trova, effettua la stessa verifica negli anni successivi alla richiesta. Si tratta di eventi eccezionali, spesso frutto di qualche errore nel salvataggio del rinnovo dell'iscrizione da un anno all'altro. La scelta di dragare le iscrizioni antecedenti e successive all'annata di interesse è stata ponderata, anche considerando il lasso di tempo relativamente limitato coperto dai dati. Da un lato, la soluzione consente di attivare correttamente i filtri e i parametri relativi all'età dell'utente, calcolati dinamicamente confrontando anno della richiesta e anno di nascita, dato che rimane invariato, salvo errori di compilazione, a prescindere dall'iscrizione selezionata. Per le altre informazioni prese in considerazione, attraverso un confronto a campione tra le diverse iscrizioni di singoli utenti, le variazioni sono risultate insignificanti, facendo presupporre, dunque, che anche nell'annata mancante i dati sarebbero stati molto simili. Va aggiunto, a scanso di equivoci, che la procedura illustrata è stata applicata solo nel caso di assenza dell'iscrizione e non per i singoli campi mancanti di un'iscrizione valida, che vengono conteggiati con la dicitura 'non dichiarato'.

Rispetto a questa procedura si riscontrano due eccezioni. Innanzitutto, esistono diversi utenti con uno specifico ID per i quali non è presente una scheda iscrizione. Si tratta, in larga parte, di personale dell'istituto, per il quale l'iscrizione annuale non era obbligatoria, che effettuò richieste per terzi o attività istituzionali. Sono state individuate, inoltre, 12.819 richieste dell'Archivio di Stato di Firenze, su un totale di 731.914, prive di un ID utente, circostanza dovuta al recupero in Archimatic di dati pregressi. Nel calcolo del numero e della percentuale di richieste, entrambe le fattispecie figurano sotto la voce 'senza iscrizione'.

Per comprendere il potenziale interesse delle metriche sviluppate, basti considerare i dati relativi all'impatto degli utenti stranieri sull'Archivio di Stato di Venezia, ricavabili applicando il filtro 'provenienza' rispetto al numero di richieste (Fig. 5). Si può notare, ad esempio, che i pezzi richiesti dall'utenza estera, terminato il periodo pandemico, hanno subito un'impennata, riportandosi addirittura ai livelli del 2015, con un trend positivo in controtendenza rispetto a quello complessivo dell'istituto veneziano e, più in generale, della maggior parte degli Archivi di Stato, che hanno faticato a riprendersi dalle chiusure del 2020-2021.¹²

¹² Sulla base dei dati statistici raccolti dal Ministero della cultura, nel 2022 gli istituti hanno dichiarato 147.820 presenze, salite a 154.447 nel 2023, contro le 257.312 del 2019. I pezzi consultati sono stati 485.698 (2022), 553.458 (2023) e 819.194 (2019).

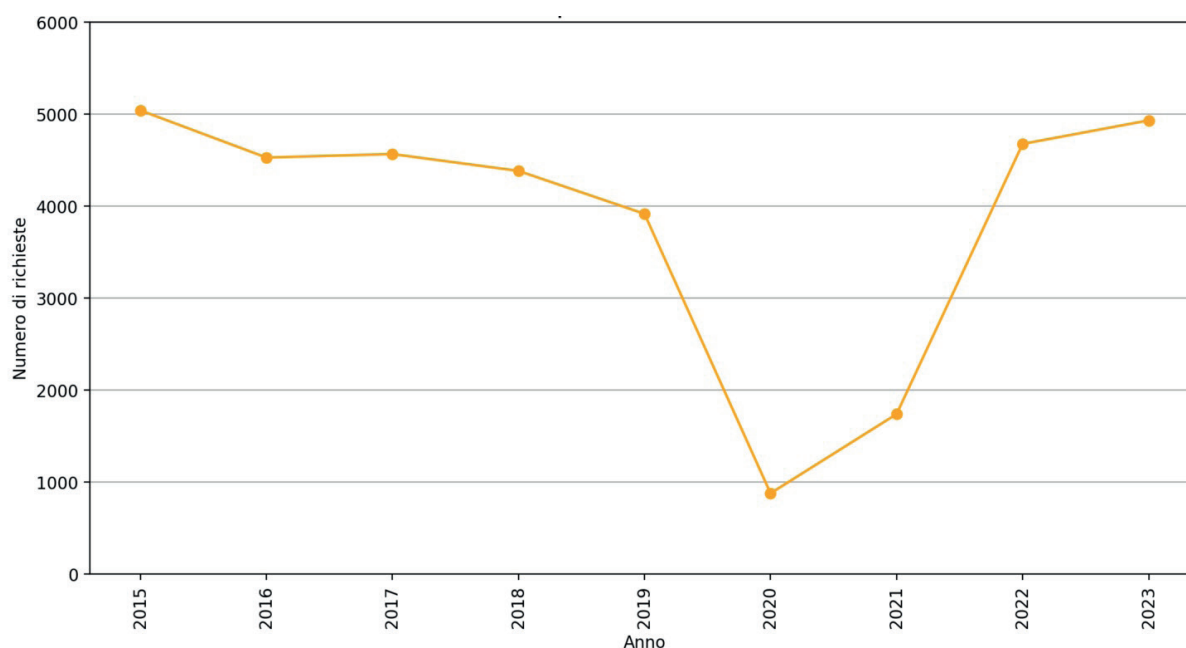


Figura 5. Grafico dei pezzi richiesti dagli utenti stranieri dell'Archivio di Stato di Venezia per il periodo 2015-2023

Utenti unici

Accanto al numero delle richieste, eventualmente filtrato, viene riportato il numero di utenti unici che le hanno effettuate, accompagnato, nel solo caso di Firenze, da quello delle richieste senza utente.¹³ Il valore degli utenti unici, vale la pena sottolinearlo, varia a seconda del periodo di tempo considerato: se in un biennio 1.000 pezzi sono stati consultati da 100 utenti unici, la somma dei pezzi consultati nel primo e nel secondo anno ovviamente sarà uguale al totale, mentre quella degli utenti unici potrebbe essere superiore, in quanto le persone che hanno consultato il materiale in entrambe le annate vengono conteggiate due volte. Anche nel caso degli utenti unici, MAI restituisce le percentuali a partire dai parametri visti in precedenza, ma in questo caso il calcolo viene effettuato esclusivamente sulle annate e non per l'intero periodo preso in esame, non essendo possibile stabilire un criterio oggettivo nella scelta di un unico valore, nel caso in cui un determinato utente presenti iscrizioni con dati differenti durante quel lasso di tempo. Un'ulteriore avvertenza riguarda le richieste fiorentine senza utente, che evidentemente non possono essere prese in considerazione nel calcolo delle percentuali degli utenti unici.

In base a quanto detto in precedenza sull'alto numero di pezzi richiesti da utenti stranieri all'Archivio di Stato di Venezia, non stupisce che anche le percentuali di utenti unici relative al parametro 'provenienza' mostrino chiaramente la vocazione internazionale dell'istituto: nell'annata 2023 gli utenti stranieri residenti all'estero hanno raggiunto addirittura il 24,7% del totale, percentuale che aumenta di alcuni punti considerando gli italiani espatriati o, in alternativa, i cittadini stranieri residenti in Italia, con entrambe le categorie che coprono il 3,6% dell'utenza (Fig. 6).

¹³ Per 'utente unico' si intende un individuo che ha effettuato almeno una consultazione nel periodo considerato e viene contato una sola volta, a prescindere dal numero di pezzi consultati.

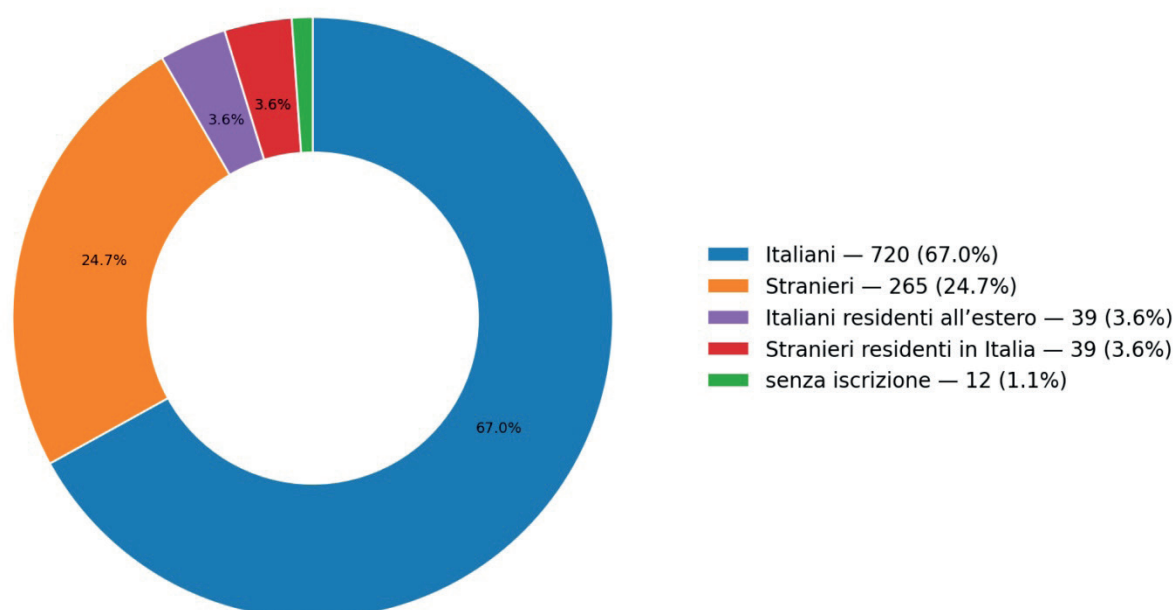


Figura 6. Grafico delle percentuali degli utenti unici dell'Archivio di Stato di Venezia divisi per provenienza per l'anno 2023

5. Prospettive future

Lo sviluppo di uno strumento per analizzare le misurazioni dei servizi erogati dagli archivi e restituire le relative metriche, per ora frutto di uno studio condotto attraverso un proficuo confronto con Francesca Fiori e Monica Del Rio, referenti dell'Archivio di Stato di Firenze e dell'Archivio di Stato di Venezia, rappresenta solo un passaggio di un processo che potrebbe dotare gli istituti di uno strumento realmente efficace nella pianificazione delle proprie strategie e nell'impiego delle risorse a disposizione. Oltre a testare e adattare le procedure illustrate sinora ad altri casi di studio, con la possibilità di servirsi di MAI per l'analisi dei dati storici di altri Archivi di Stato, il progetto intende proporsi come punto di partenza per analizzare e migliorare quanto più possibile la raccolta e l'analisi futura delle informazioni, superando alcuni dei limiti riscontrati. In particolare, sono stati individuati due aspetti da approfondire: la standardizzazione dei vocabolari utilizzati per classificare l'utenza e l'individuazione di ulteriori misurazioni e metriche.

Standardizzazione dei vocabolari

Per ovviare ad alcune delle criticità riscontrate nell'analisi dei dati pregressi, sarebbe opportuno standardizzare i vocabolari, facendo riferimento a soggetti, liste di autorità o specifici standard, nazionali e internazionali. In particolare, andrebbero sottoposte a una revisione puntuale le liste relative a 'professione', 'cittadinanza/nazionalità', 'residenza', 'ambito di ricerca', 'ambito cronologico', 'motivo della richiesta', ripristinando la voce 'titolo di studio'.

A livello internazionale, si potrebbe far riferimento alle authority dei *Library of Congress Demographic Group Terms* - LCDGT, selezionando un set di valori ricavati attraverso i seguenti filtri:

‘occupation’ (761 occorrenze), ‘nationality’ (655); ‘age’ (17); ‘educational’ (41).¹⁴ In alternativa, optando per standard specifici, la classificazione delle professioni, dei titoli di studio e degli ambiti disciplinari potrebbero basarsi rispettivamente sull’*International Standard Classification of Occupations - ISCO-08*,¹⁵ sull’*International Standard Classification of Education - ISCED 2011*¹⁶ e sui *Fields of Science and Technology (FOS)* pubblicati dall’Organisation for Economic Co-operation and Development - OECD nel 2002, che nella versione attuale sono alla base anche della tassonomia definita dall’*European Science Vocabulary - EuroSciVoc*.¹⁷ Per la codifica geografica esistono numerosi strumenti, a cominciare dallo standard ISO 3166 - *Codes for the representation of names of countries and their subdivisions*,¹⁸ mentre per l’ambito cronologico si potrebbe far riferimento a *PeriodO*, repertorio di periodi storici in formato LOD.¹⁹

A livello nazionale, per varie classificazioni si potrebbero utilizzare le voci legate ad alcuni nodi del *Nuovo soggettario* della Biblioteca nazionale centrale di Firenze, come ‘Persone secondo l’attività’ o ‘Discipline’.²⁰ Punti di riferimento più specifici si ritrovano nelle classificazioni adottate dall’ISTAT, tra le quali: la *Classificazione delle professioni CP2021*,²¹ la *Mappa aggiornata dei percorsi di istruzione e dei titoli di studio CLAIST 2025*, definita in collaborazione con un Comitato interistituzionale di esperti del settore;²² i *Codici delle unità amministrative*²³ e i *Codici delle unità territoriali estere*,²⁴ attraverso i quali è possibile monitorare le variazioni nelle circoscrizioni territoriali italiane e dei paesi stranieri. Per gli ambiti di ricerca, esistono i settori scientifico-disciplinari previsti dall’art. 15 della legge 30 dicembre 2010, n. 240 sul sistema universitario.²⁵ L’ambito cronologico, infine, potrebbe basarsi sul *Thesaurus per le indicazioni cronologiche* dell’Istituto Centrale per il Catalogo e la Documentazione - ICCD.²⁶

Individuazione di ulteriori misurazioni e metriche

La possibilità di effettuare misurazioni puntuali e automatizzate del numero di presenze giornaliere degli utenti o dei tempi di consultazione del materiale, così come di altri valori ‘critici’, dipenderà dalla disponibilità di infrastrutture hardware e software adeguate, ma anche, per non dire soprattutto, dalla volontà degli istituti di sensibilizzare e formare l’utenza e il personale addetto ai servizi di sala studio sull’uso di strumenti e procedure considerati spesso un inutile aggravio nello svolgimento del lavoro quotidiano. Dipenderà da questo cambio di mentalità anche l’uso consapevole delle numerose funzionalità già presenti in Sala Studio, piattaforma che consente non solo

¹⁴ <https://id.loc.gov/search/?q=cs:http://id.loc.gov/authorities/demographicTerms>.

¹⁵ <https://isco-ilo.netlify.app/en/isco-08/>.

¹⁶ <https://uis.unesco.org/en/topic/international-standard-classification-education-isced/>.

¹⁷ <https://technoportal.hevs.ch/ontologies/EUROSCIVOC>.

¹⁸ <https://www.iso.org/iso-3166-country-codes.html>.

¹⁹ <https://perio.do>.

²⁰ <https://thes.bncf.firenze.sbn.it/>.

²¹ <https://www.istat.it/classificazione/classificazione-delle-professioni/>.

²² <https://www.istat.it/classificazione/classificazione-dei-titoli-di-studio-italiani/>.

²³ <https://www.istat.it/classificazione/codici-dei-comuni-delle-province-e-delle-regioni/>.

²⁴ <https://www.istat.it/classificazione/classificazione-degli-stati-esteri/>.

²⁵ <https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:legge:2010-12-30:240>.

²⁶ <http://www.iccd.beniculturali.it/it/ricercanormative/191/thesaurus-per-le-indicazioni-cronologiche>.

di gestire le normali richieste di consultazione dei pezzi archivistici, ma anche una serie di servizi che vanno dal rilascio di riproduzioni digitali o copie conformi dei documenti, alla prenotazione di una consulenza in sede, passando per le ricerche da remoto o la partecipazione a un evento. Al netto di questo salto di qualità e dell'ampia gamma di informazioni che si potrebbero raccogliere, un'ulteriore prospettiva di sicuro interesse è legata alla ricchezza e alla qualità delle descrizioni del patrimonio archivistico italiano che saranno progressivamente migrate e prodotte nel nuovo Sistema Informativo Archivistico sviluppato dall'Istituto Centrale per gli Archivi. La possibilità di associare a ogni richiesta una serie di metadati derivati dalle schede descrittive del complesso archivistico interessato o, addirittura, della singola unità archivistica o documentaria, aumenterà a dismisura i valori passibili di misurazione, consentendo un'analisi dell'utilizzo dei fondi basata su una serie di parametri, come l'epoca storica del materiale consultato, da incrociare con le notizie fornite dagli utenti in fase di iscrizione. In questo caso, l'efficacia delle rilevazioni non dipenderà più dalla qualità dei form di iscrizione o dalla precisione degli utenti nel compilarli, ma dalla produzione di descrizioni archivistiche di qualità, che dovrebbe essere, vale la pena ricordarlo, una delle mansioni, o per meglio dire ambizioni, dell'archivista.

Riferimenti bibliografici*

- Chiarelli, Debora. 2024. *Il valore economico degli archivi*. Roma: Ministero della cultura.
- Duff, Wendy M., Elizabeth Yakel, Helen R. Tibbo, Joan M. Cherry, Aprille Mckay, Magia G. Krause, e Rebecca Sheffield. 2010. "The development, testing, and evaluation of the archival metrics toolkit." *The American Archivist* 73 (2): 569-99. <https://doi.org/10.17723/aarc.73.2.00101k28200838k4>.
- Faggiolani, Chiara. 2019. *Conoscere gli utenti per comunicare la biblioteca: Il potere delle parole per misurare l'impatto*. Milano: Editrice Bibliografica.
- Feliciati, Pierluigi. 2012. "Ask the users, il valore aggiunto della valutazione dei sistemi informativi culturali on line coinvolgendo gli utenti: il caso del progetto Una Città per gli Archivi." *Il capitale culturale. Studies on the Value of Cultural Heritage* 5: 129-44. <http://dx.doi.org/10.13138/2039-2362/483>.
- Feliciati, Pierluigi. 2016. "L'usabilità degli ambienti bibliotecari e archivistici digitali come requisito di qualità: contesto, modelli e strumenti di valutazione." *JLIS.it* 7 (1): 113-30. <http://dx.doi.org/10.4403/jlis.it-11512>.
- Feliciati, Pierluigi. 2023. "Gli utenti archivistici e gli strumenti AI per l'orientamento e l'accesso: uno studio nell'ambito del progetto InterPARES Trust AI." *JLIS.it* 14 (3): 117-28. <https://doi.org/10.36253/jlis.it-558>.
- Feliciati, Pierluigi, e Giorgia Di Marcantonio. 2023. "Users' approaches and behaviors in accessing records and archives in the perspective of AI: a global user study Phase 1 - Italian study, Vancouver, Interpares Trust AI". <https://doi.org/10.5281/zenodo.7852021>.
- Gardini, Stefano, e Mauro Giacomini. 2019. "Venticinque anni di consumi e produzioni culturali: aspetti quantitativi e spunti qualitativi dal database della sala di studio dell'Archivio di Stato di Genova (1991-2016)." In *Ianuensis non nascitur sed fit. Studi per Dino Puncuh*, 619-68. Genova: Società Ligure di Storia Patria.
- Gardini, Stefano. "Utenti e usi dell'archivio: prospettive storiche e profili tipologici dal caso dell'Archivio di Stato di Genova (1883- 2015)". Tesi di dottorato, Università di Roma La Sapienza, 2020.
- Hawk, Amanda K., e Melanie Griffin. 2022. "Adherence to Standardized Measures and Metrics for Public Services in U.S.-Based Special Collections Libraries and Archival Institutions." *Journal of library administration* 62 (5): 673–88. <https://doi.org/10.1080/01930826.2022.2083443>.
- Lanzini, Marco. 2022. "Un'eco lontana da cogliere e interpretare: proposte per un'analisi statistica dell'utenza degli archivi attraverso il caso di studio dell'Archivio di Stato di Milano." *Nuovi Annali Della Scuola Speciale per Archivisti e Bibliotecari* 36: 371-407.
- NCA. PSQG (National Council on Archives. Public Services Quality Group). 2008. *A Standard for access for archives*. https://static1.squarespace.com/static/60773266d31a1f2f300e02ef/t/61b-855baf9e15d5909a1060a/1639470522963/access_standard_2008.pdf.

* Ultima data di consultazione dei siti web: 9 dicembre 2025.

Pickford, Chris. 2002. "Archives: a statistical overview." *Cultural Trends* 12 (48): 1-36. <https://doi.org/10.1080/09548960209390339>.

Solimine, Giovanni, Prefazione a *Conoscere gli utenti per comunicare la biblioteca: Il potere delle parole per misurare l'impatto*, di Chiara Faggiolani (Milano: Editrice Bibliografica, 2019).

Wikidata and authority files: reconciliation and cooperation procedures

Camillo Carlo Pellizzari di San Girolamo^(a)

a) Scuola Normale Superiore, <https://orcid.org/0000-0003-2699-1693>

Contact: Camillo Carlo Pellizzari di San Girolamo, camillo.pellizzaridisangirolamo@sns.it

Received: 06 August 2025; **Accepted:** 19 February 2026; **First Published:** 15 May 2026

ABSTRACT

This article aims to providing a comprehensive overview of the possible interactions between Wikidata and authority files, which can be classified into two main categories: firstly, the methods of reconciliation between Wikidata items and authority records; secondly, the ways in which the Wikidata community can improve Wikidata items using the authority records to which they link, and in which the cataloguers can improve their authority records using the Wikidata items that link to them, in a collaborative perspective. The purpose of this overview is optimising existing workflows and encouraging to establish new ones, so as to foster the reciprocal improvement of Wikidata's and authority files' data. The last part of the article describes a few major collaborations between Wikidata and authority files of national and international relevance, and the main attempts to use Wikibase (Wikidata's software) as a platform to store authority files.

KEYWORDS

Wikidata; Authority control; Entity reconciliation; AuthorityBox; Wikibase.

1. Introduction

1.1 Historical background

The cooperation between Wikidata (from now onwards WD) and authority files (from now onwards AFs, or singular AF) started in 2013: on March 6, the first properties with datatype “string” (afterwards converted to “external-id”) were created in WD, specifically P212 for ISBN-13, P213 for ISNI, and P214 for VIAF.

In the same year the first significant reconciliation of authority records (from now onwards ARs, or singular AR) to WD was performed as part of a project by Maximilian Klein, Wikipedian in Residence at OCLC, and Andrew Gray, Wikipedian in Residence at the British Library: VIAFbot, first proposed in June 2012, was initially programmed to import to English Wikipedia articles the links to the VIAF clusters that already linked to them, but since March 2013 it was adapted to operate on WD (Klein and Kyrios 2013).

In 2016 the International Federation of Library Associations and Institutions (IFLA) endorsed a white paper acknowledging that «Wikidata is becoming an important tool for synchronizing across identifiers like Virtual International Authority File (VIAF) and ORCID identifiers» (Bartholmei et al. 2016, 10).

In 2019 the Association of Research Libraries (ARL) published a white paper regarding WD, in which a paragraph dealt with «Authority Data and Using Wikidata as a Linking Hub». The white paper encouraged the collaboration between WD and AFs, stating that «Wikidata has the potential to be an important part of the linked data ecosystem for authorities» (Association of Research Libraries 2019, 29). AFs are crucial for WD to connect to the bibliographic resources authored by or containing information about the entities it describes, and they are also valuable sources for the structured data about these entities (e.g. birth and death dates). At the same time, WD can provide to libraries persistent identifiers (PIDs) and a vast amount of data (both factual data and external IDs) that they can use to enrich their ARs and, more generally, their catalogues, providing more context to the users.

In the same period many groups were created to explore the use of WD in GLAM institutions, and especially in libraries, including the LD4 Wikidata Affinity Group¹ in 2019 and the Gruppo Wikidata per Musei, Archivi e Biblioteche (GWMAB)² in 2020; a Wikidata Working Group in IFLA has been active from January 2020 to December 2022;³ in 2020-1 the Program for Cooperative Cataloging (PCC) sponsored a PCC Wikidata Pilot⁴ to guide librarians in experimenting the use of WD (cf. Zhang et al. 2023). The above list is not exhaustive and focuses on projects having authority control among their main focuses (cf. Boccone 2022).

Systematic reviews of WD’s use by libraries (Tharani 2021) and, more broadly, by GLAM (Galleries, Libraries, Archives, and Museums) institutions (Candela et al. 2024) both outlined a growing interest for WD. In the specific case of libraries, «Integration and interoperability of authority

¹ https://www.wikidata.org/wiki/Wikidata:WikiProject_LD4_Wikidata_Affinity_Group.

² https://www.wikidata.org/wiki/Wikidata:Gruppo_Wikidata_per_Musei_Archivi_e_Biblioteche.

³ https://www.wikidata.org/wiki/Wikidata:IFLA_Wikidata_Working_Group.

⁴ https://www.wikidata.org/wiki/Wikidata:WikiProject_PCC_Wikidata_Pilot.

data using Wikidata emerged as the most popular application area based on the number of articles coded for this category» (Tharani 2021, 3).

The importance of WD in authority control has been further highlighted through comparison with VIAF (Bianchini et al. 2021) and in the perspective of Universal Bibliographic Control (Sardo and Bianchini 2022); WD also allows to extend the concept of authority control to researchers, establishing a connection between library catalogues and research repositories (Bargioni et al. 2023). In many digital humanities projects WD itself is currently used as an AF (Zhao 2023, 868); the procedures for creating personal WD items and personal ARs have also been compared, tracing similarities and differences (Chen and Yuxuan 2024).

Some major collaborations between AFs and WD are described in chapter 5.1.

1.2 Statistics about the present situation

In WD, library AFs constitute a small but significant group of properties: as of now the properties that are “instance of” (P31) “Wikidata property for authority control (libraries)” (Q96776953) or of one its subclasses, which include mainly “Wikidata property for authority control by VIAF member” (Q55586529), are 220 in total,⁵ out of 9,569 external-id properties.⁶ These properties are used as value in 22.8 M statements⁷ contained in 5.6 M unique items,⁸ the majority of them (4.2 M) being personal items (i.e. having P31 = Q5, “instance of = human”);⁹ restricting the queries to VIAF members only, they are used as value in 18.7 M statements¹⁰ contained in 4.6 M unique items¹¹ (including 3.7 M personal items;¹² as a comparison, in 2020 VIAF contained 22.1 M personal clusters¹³).

⁵ <https://qlever.dev/wikidata/LI6Z15>.

⁶ <https://qlever.dev/wikidata/4avd1b>.

⁷ <https://qlever.dev/wikidata/wR9jD3>.

⁸ <https://qlever.dev/wikidata/INhZmi>.

⁹ <https://qlever.dev/wikidata/ZwukeP>.

¹⁰ <https://qlever.dev/wikidata/DrpXS>.

¹¹ <https://qlever.dev/wikidata/oTzGAR>.

¹² <https://qlever.dev/wikidata/k0ZRWv>.

¹³ https://catalogo.pusc.it/beyond_viaf/.

Among the aforementioned 220 properties, the ten used as values in the highest number of distinct items are:

Table 1. The ten properties linking to library AFs used as values in the highest number of distinct items. The number of distinct items using the properties comes from a SPARQL query,¹⁴ whilst the number of users having added them at least once as main statement (from their creation and in the month of June 2025) comes from Wikidata Navel Gazer¹⁵

property	VIAF code	items (2024-10-12)	items (2025-07-21)	users (total)	users (June 2025)
P214	(VIAF)	3,444,466	3,649,450	10,439	615
P227	DNB	1,843,039	2,426,148	5,314	318
P213	ISNI	1,699,662	2,216,034	6,840	424
P244	LC	1,635,913	1,686,829	6,750	427
P269	SUDOC	982,160	1,007,193	3,302	255
P8189	J9U	823,439	835,048	1,304	127
P691	NKC	783,355	813,295	1,420	70
P268	BNF	701,258	723,772	4,019	298
P1207	NUKAT	632,937	687,557	1,645	155
P1006	NTA	559,425	576,486	1,649	152

2. Reconciliation

The premise for a comparison between the data in WD and the data in an AF is that WD items and ARs should have been reconciled firstly.

2.1 Interlinking

WD items can link to ARs through apposite properties: if an AF is publicly accessible and assigns a URI to each AR, an external-id property can be proposed¹⁶ and, after a community discussion of minimum 7 days, it is created if there is consensus (i.e. if no objection is raised, or when objections have been adequately answered).

ARs can link to WD items through the 024 field (“other standard identifier”) in MARC 21 format or the 017 field (“other identifier”) in UNIMARC format; in both cases \$a is used for the ID and \$2 for the source indication “Wikidata”, or a chosen abbreviation.

In principle, the correspondence between ARs and WD items should be biunivocal (1:1), but exceptions are possible, e.g. in the case of pseudonyms: WD usually has one item for the author as a person and for all their pseudonyms (but exceptions are made wherever one Wikimedia project has

¹⁴ <https://qlever.dev/wikidata/Uupogc>.

¹⁵ <https://bambots.brucemyers.com/NavelGazer.php>.

¹⁶ https://www.wikidata.org/wiki/Wikidata:Property_proposal.

distinct pages for the person and for some of their pseudonyms); similarly, certain cataloguing rules prescribe having only one AR for each physical person (e.g. REICAT, used by SBN), whilst others (e.g. RDA) prescribe creating distinct ARs for the person and for each of their pseudonyms if the person has used both their real name and one or more pseudonyms in their works (De Monaco 2025, 7–21; Chou 2025). In cases like this the WD property can apply to the “single-value constraint” (Q19474404) the qualifier “separator” (P4155), which allows to define specific qualifier-properties to be used as separators, in order to mark manually the cases in which the non-biunivocal correspondence is intentional and thus does not constitute a mistake in WD and/or in the AF.¹⁷

Apart from exceptions, the other cases of non-biunivocal correspondence between ARs and WD items are usually identification issues on one or both sides; undertaking the reconciliation of an AF with WD is usually a perfect occasion to find them (De Monaco 2025, 49–51; De Monaco and Pellizzari di San Girolamo 2025, file ‘Tracciamento errori riscontrati in SBN.xlsx’).

2.2 Available tools

WD items and ARs can be reconciled manually, adding in an AR a link to a WD item through the 024 MARC 21 field or the 017 UNIMARC field and adding in a WD item the link to an AR through its apposite property. This could be a standard practice for the cataloguers while manually creating new ARs or editing existing ARs; if the link is manually created in one of the two ways, an automatic system can be established to periodically add the reciprocal link (see e.g. NKC in chapter 5.1).

To import to WD a pre-existing list of matches between WD items and ARs, QuickStatements (from now onwards QS) is the most convenient choice;¹⁸ the tool can be used by autoconfirmed users¹⁹ and allows to enter these matches with a CSV syntax at a rate of usually many tens per minute.

Matching in a completely manual way with WD an AF consisting of tens or hundreds of thousands of ARs is obviously not feasible. For this reason, in order to match old ARs with WD items, two reconciliation tools can be used: Mix’n’match (from now onwards MnM)²⁰ and OpenRefine (from now onwards OR)²¹. Both these tools, given a set of entries, try to match them with WD, suggesting one or more matches whenever possible for each entry and allowing to create new WD items for the others; the user can always revise these judgements, confirming or refusing them.

Apart from these similarities, MnM and OR have many differences. MnM is an online platform used to reconcile WD items to other (usually online) datasets: the datasets are uploaded as catalogs,²²

¹⁷ See e.g. <https://www.wikidata.org/wiki/Property:P1368#P2302>.

¹⁸ <https://quickstatements.toolforge.org/> (manual: <https://www.wikidata.org/wiki/Help:QuickStatements>). Since 2024, a new version of QS with additional features is also available, QS 3.0 (<https://qs-dev.toolforge.org/>; manual: https://meta.wikimedia.org/wiki/QuickStatements_3.0/Documentation).

¹⁹ https://www.wikidata.org/wiki/Wikidata:Autoconfirmed_users.

²⁰ <https://mix-n-match.toolforge.org/> (the manual, <https://meta.wikimedia.org/wiki/Mix%27n%27match/Manual>, is significantly out-of-date as of now).

²¹ Software downloadable from <https://openrefine.org/>, also containing documentation; a manual in WD is also available (<https://www.wikidata.org/wiki/Wikidata:Tools/OpenRefine>).

²² E.g. <https://mix-n-match.toolforge.org/#/catalog/4129>.

which can be either scraped through an apposite MnM function, or uploaded through a CSV/TSV file or a public Google Spreadsheet; MnM catalogs are public, so all WD users can cooperate to their reconciliation (MnM can also be used directly inside WD items thanks to an apposite gadget²³); a significant strength of MnM is containing thousands of catalogs, so it is a very good tool to gain an understanding of e.g. conflicting homonyms among many AFs and other sources, and it also allows to create new items combining data from many sources. OR is a software, allowing to clean-up datasets and to reconcile them with WD items; its reconciliation functions and its capability of creating high-quality new items (e.g. with statements containing qualifiers and elaborate references) are more sophisticated than those of MnM: however, in OR a single user works on a single dataset, without the possibility of crossing one's own dataset with many other catalogs as in MnM. Consequently, MnM tends to be more effective for projects where it is useful to have a reconciliation performed on big amounts of data by potentially many users in a long span of time, and in cases where a majority of the entries already exist as WD items, whilst OR is surely more suitable for smaller projects managed by one user, and in cases requiring the majority of entries to be created as new WD items.

3. Identification issues in authority records and Wikidata items

When reconciling ARs and WD items, their quality plays a significant role in making the process easy or difficult, sometimes impossible. In assessing the quality of an entry (AR or WD item), three parameters are taken into account:

1. identification: the entry is clearly self-identified,²⁴ not self-identified²⁵ or confusedly self-identified;²⁶
2. incoming links: the links to the entry are only appropriate, or partly inappropriate (note: the distinction appropriate/inappropriate is based on the entry being clearly self-identified; if it is not self-identified, the check should instead establish if the incoming links relate to one entity or 2(+) entities);
3. duplicates: the entry has no duplicates, or 1(+) duplicates.

Parameters 1 and 3 apply to both ARs and WD items; parameter 2 usually applies to ARs, since an AR usually exists if it is connected to at least one bibliographic record (from now onwards BR, or plural BRs), but rarely to WD items, since a WD item can exist without any incoming link from other WD items.

On the basis of these parameters, the following table (Table 2) classifies the identification issues that can affect ARs and WD items.

²³ https://www.wikidata.org/wiki/User:Magnus_Manske/mixnmatch_gadget.js.

²⁴ The entry contains sufficient information to make the described entity clearly identifiable: usually the dates of birth and death, or the dates of activity, and the occupation(s).

²⁵ The entry contains insufficient information to make the described entity clearly identifiable (in the most common case, it contains only a name).

²⁶ The entry contains incoherent information, with some data pointing to an entity A and some other to an entity B, so that it is unclear if the entry describes A or B.

Table 2. Table of possible identification issues in ARs and WD items

Situation code and name	Description in ARs	Description in WD items
S0: “no issue”	clearly self-identified AR with all the appropriate BRs connected	clearly self-identified WD item with all the appropriate links from other WD items
S1: “wrong connections”	clearly self-identified AR but with 1(+) inappropriate BR(s) connected	clearly self-identified WD item but with 1(+) inappropriate link(s) from other WD items
S2: “weakly identified”	not self-identified AR; identification rests upon the connected BR(s), which relate to only one entity	(theoretically impossible in WD: according to the notability criteria, ²⁷ items must always be “clearly identifiable” in themselves; if they are not, they should be fixed or deleted)
S3: “unidentified”	not self-identified AR; identification cannot rest upon the connected BR(s), since they relate to 2(+) entities	
S4: “conflated”	confusedly self-identified AR	confusedly self-identified WD item
SD: “duplicates”	2(+) ARs regarding the same entity	2(+) WD items regarding the same entity
* SD0: “duplicates, no (other) issues”	2(+) clearly self-identified ARs regarding the same entity with no inappropriate BRs connected	2(+) clearly self-identified WD items regarding the same entity with no inappropriate links from other WD items
* SD1: “duplicates, with wrong connections”	2(+) clearly self-identified ARs regarding the same entity with 1(+) inappropriate BR(s) connected	2(+) clearly self-identified WD items regarding the same entity with 1(+) inappropriate link(s) from other WD items
* SD2: “duplicates, weakly identified”	2(+) not self-identified ARs which, according to their identification based on the connected BR(s), relate to only one entity	<i>same as for S2 and S3</i>

In WD items, given that most of them do not have any links from other WD items, S1 is rare, as well as S2 and S3 (which are theoretically impossible, but may happen when items are based on sources which contain very few identifying details about them). So basically in WD the focus of the user can be mainly on the item itself, since the incoming links are mostly absent.

Conversely, ARs usually have at least one connected BR, so their identification is a twofold issue, focusing both on the AR itself and on its connected BR(s). S0 is the standard to which all ARs should point; an S0-AR degrades to S1 when 1(+) inappropriate BR(s) are connected to it, but since it contains good identifying information the inappropriate BR(s) can be easily found and removed, restoring the AR to S0. The rare situation S4 happens when identifying information is poorly assembled and mixes 2(+) entities which are in fact distinct, making the AR completely unusable. S3 is the status of a catalogue where authority control is not implemented, or poorly applied. S2 stays in the middle, i.e. a catalogue where BR(s) are correctly attributed but ARs do not contain

²⁷ <https://www.wikidata.org/wiki/Wikidata:Notability>.

identifying information; S2 is not a mistake in itself, but can easily degrade into SD2 and/or S3 when new BR(s) are wrongly attributed due to insufficient guidance given by AR(s) themselves.

4. Five ways of reciprocal improvement

As said, both WD items and ARs can contain mistakes (i.e. cases S1, S3, S4, SD) and incomplete entries (i.e. case S2), and these mistakes are likely to emerge when someone tries to reconcile WD with an AF. However, while a random WD user can edit directly WD items to solve problems of identification, improving them to S0, the same is not true for AFs, where only the employees of the institution(s) creating it, or users authorized by the institution(s), can make edits, so their mistakes cannot be solved directly by most WD users (cf. Pellizzari di San Girolamo 2024a, para. 5.2).

This chapter explores how the collaboration between WD and AFs (i.e. between WD users and AFs' editors) can lead to increasing the percentage of both WD items and ARs reaching S0. WD, while allowing AFs to reduce their amount of mistakes, also constitutes for them a good way to increase their incoming and outgoing links (both to other AFs and to different kinds of resources), thus improving their position in the Linked Data Web.

4.1 Data round-tripping

WD users, noticing mistakes in ARs, try reporting them, in a process known in WD as “data round-tripping”,²⁸ usually through a webform or an e-mail provided by the institution(s) managing the AF (and indicated in the property itself through P10923, “error-report URL or e-mail”), or contacting directly another WD user with editing access to it;²⁹ unfortunately, for some AFs there are no known mistake reporting methods (Pellizzari di San Girolamo 2025).

Table 3. Table of the usual management strategies of ARs' identification issues in WD items

Situation in AR	Strategy in WD: add the AR ID(s)
S0: “no issue”	
S1: “wrong connections”	and try reporting the wrong BR(s)
S2: “weakly identified”	
S3: “unidentified”	with deprecated rank and the qualifier P2241 (“reason for deprecated rank”) = Q14946528 (“conflation”), and try reporting the conflation
S4: “conflated”	
SD0: “duplicates, no (other) issues”	
SD1: “duplicates, with wrong connections”	and try reporting the wrong BR(s)
SD2: “duplicates, weakly identified”	

²⁸ https://www.wikidata.org/wiki/Wikidata:Data_round-tripping.

²⁹ https://www.wikidata.org/wiki/Wikidata:WikiProject_Authority_control/Error_reporting_procedures.

From WD's perspective (shown in Table 3), adding AR ID(s) is always useful, also when they do not have S0 quality, because it both makes their mistakes more evident and it helps avoiding possible mistakes by other WD users. Specifically, there is no need to send manual reports for SD0 and SD2, as these cases can be listed automatically (see chapter 4.2). S1 and SD1 do not need high-priority mistake reports, but their inappropriate BR(s) should anyway be reported as they are likely to cause confusion in other AFs and especially to lead to VIAF conflations. Manual reports are absolutely necessary only for S3 and S4, since these ARs cannot be linked from WD unless properly fixed; until they are fixed, they must be deprecated to avoid other WD users considering them reliable ARs.

4.2 Constraint violations and mismatches

WD uses constraints to assess potentially wrong statement values,³⁰ a constraint violation (from now onwards CV, or plural CVs) is visualized through an icon that warns users to check strange values and correct them if needed; lists of CVs can be obtained through SPARQL queries automatically generated by the Property documentation template³¹ at the top of the talk page of each property.³² In order to check identification issues of WD items and ARs, two types of CVs are particularly relevant: unique-value CVs, i.e. the same external ID linked from 2(+) WD items, and single-value CVs, i.e. the same WD item linking to 2(+) external IDs in the same database.

Table 4. Table indicating how unique-value and single-value CVs can be used to find duplications and conflations in WD items and ARs (adapted from Pellizzari di San Girolamo 2023, 14)

	WD items	ARs
duplication (= SD)	unique-value CV	single-value CV
conflation (= S3/S4)	unique-value CV single-value CV	unique-value CV single-value CV

As Table 4 shows, unique-value and single-value CVs always show mistakes in WD items, ARs or both (apart from the few cases of normal non-biunivocal correspondence between them; see chapter 2.1); when mistakes are solved, the cases disappear from the lists of CVs. However, as WD users can usually solve only mistakes in WD items, the cases depending from mistakes in ARs tend to remain in the lists of CVs for a long time and to make them more difficult to use; in particular, the high amount of SD cases in ARs tends to flood the lists of single-value CVs, making very difficult to find through them conflations in WD items and in ARs.

Since the lists of CVs are the most powerful tool for both WD users and AFs' editors to discover identification issues, monitoring and emptying them constantly is the best way to keep them readable and usable; ideally, AFs' editors should be involved as WD users in fixing the mistakes shown by the lists of CVs both in WD items and in their ARs. In 2024 the WikiProject Disambiguation

³⁰ https://www.wikidata.org/wiki/Help:Property_constraints_portal.

³¹ https://www.wikidata.org/wiki/Template:Property_documentation.

³² E.g. https://www.wikidata.org/wiki/Property_talk:P396.

has been created in WD to connect AFs' editors active in WD and to make easy reporting to them the most complex cases of conflations affecting multiple AFs.³³

A query for deprecated statements should be used to find previously-marked conflations; this query is also generated by the Property documentation template; as in the case of other queries, it is important to gradually normalize in WD the AR IDs once fixed, so that the query does not always show the same unsolved results.

4.3 Manual editing on Wikidata

SPARQL queries in WD can be used to find items already connected to a certain AF but with significant information missing (e.g. items without VIAF,³⁴ or without birthdate,³⁵ or without occupation,³⁶ or with conflicting birthdates,³⁷ or with unreferenced birthdate³⁸); for AFs' editors, using these queries to find and improve these items has a lower priority in comparison with the periodical check on the lists of CVs (see chapter 4.2), but can still have two important aims: firstly, making easily available good quality authority data, so that AFs' editors in other countries can use them to improve their ARs to S0; secondly, being able to make reliable statistics about the authors described in the AF (e.g. regarding their distribution by birthplaces, deathplaces, lifespan) through WD SPARQL queries (De Monaco 2025, 60–9).

4.4 Massive enrichment from Wikidata

WD items can be used as a source to import data into an AF; the data can be copied directly into the ARs and/or displayed in real time through the API, e.g. through an AuthorityBox (Bargioni 2020). Examples of both enrichments are mentioned in chapter 5.1.

Both solutions have relevant pros and cons that should be evaluated. From a technical point of view, copying data into the ARs requires being able to massively edit ARs in the backend, whilst displaying data from WD means changing the frontend visualization of ARs. This second task could be difficult or impossible when using a proprietary ILS.

Copying the data directly into the ARs has the advantage of definitively improving the quality of ARs (given that only referenced data are taken from WD), particularly improving S2 ARs to S0, so as to avoid them degrading into S3; however, this data import is static (i.e. it does not consider further updates and/or improvements made in WD items, unless repeated periodically) and adapting WD's data to the data format used in the AF might be challenging, especially if the ARs have a low data granularity (i.e. a lot of information is expressed by a textual note, instead of being segmented in different fields for dates, places, occupation(s) etc.) and thus it is necessary to concatenate the data in readable phrases.

³³ https://www.wikidata.org/wiki/Wikidata:WikiProject_Disambiguation.

³⁴ E.g. <https://w.wiki/Hy3i>.

³⁵ E.g. <https://w.wiki/Hy3k>.

³⁶ E.g. <https://w.wiki/Hy3n>.

³⁷ E.g. <https://w.wiki/Enmr>.

³⁸ E.g. <https://w.wiki/EnkA>.

Conversely, displaying data in real time has the advantage of using always up-to-date data, but in some cases it can lead to showing incoherent data (i.e. divergencies between what is written in the AR and what is displayed from the corresponding WD item).

4.5 Massive imports to Wikidata

Massively importing data from an AF to WD has basically the same advantages of manually editing WD items (cf. chapter 4.3), but on a bigger scale: the data in the AF become much more visible and reusable for other AFs, and easily analysable through WD SPARQL queries. Massive imports are possible only if the AF has released its data in CC0 license, a choice already made by GND (since 2015),³⁹ NKC (since 2022),⁴⁰ and NTA;⁴¹ otherwise, only one-by-one manual imports are legally possible.⁴² The most relevant imports of data from AFs to WD have been made by NKC (cf. chapter 5.1).

5. Some relevant collaborations

5.1 Authority files actively collaborating with Wikidata

This chapter provides brief descriptions of the most relevant ongoing collaborations between AFs and WD.

NKC (P691 in WD): the import of ARs' data into WD items, including the creation of new missing items, began in 2019, as part of a pre-existing collaboration between the National Library of Czech Republic (NLCZ) and Wikimedia Czech Republic (WMCZ); the total uses of P691 climbed from 220,732 at the end of 2018⁴³ to 264,736 at the end of 2019,⁴⁴ 646,660 at the end of 2020,⁴⁵ 1,656,882 at the end of 2022;⁴⁶ they are currently 2,118,160.⁴⁷ The reconciliation was made firstly importing pre-existing matches from Czech Wikipedia, then with semiautomatic methods: VIAF matches were tried but abandoned due to the excessive mistake percentage; the most effective semiautomatic method proved to be combining the name match with a date match (same birthdate, i.e. day of birth, or same year of birth and year of death, or same year of birth and same VIAF); in parallel, MnM is also used for manual reconciliation (Dostál 2021). Newly created ARs can be monitored in WD week by week⁴⁸ and, most importantly, a bot⁴⁹ imports on a daily basis

³⁹ https://gnd.network/Webs/gnd/EN/FAQ/faq_node.html; https://www.dnb.de/DE/Professionell/Metadatendienste/Datenbezug/Gesamtabzuege/gesamtabzuege_node.html.

⁴⁰ <https://www.nkp.cz/en/about-us/professional-activities/open-data>.

⁴¹ <https://data.bibliotheken.nl/KB/-/stories/NTA>.

⁴² https://www.wikidata.org/wiki/Wikidata:Data_donation.

⁴³ <https://www.wikidata.org/w/index.php?oldid=823585223&action=edit>.

⁴⁴ <https://www.wikidata.org/w/index.php?oldid=1087497651&action=edit>.

⁴⁵ <https://www.wikidata.org/w/index.php?oldid=1332488380&action=edit>.

⁴⁶ <https://www.wikidata.org/w/index.php?oldid=1799601120&action=edit>.

⁴⁷ <https://www.wikidata.org/w/index.php?oldid=2372675275&action=edit>.

⁴⁸ https://www.wikidata.org/wiki/Wikidata:WikiProject_Czech_Republic/New_authorities.

⁴⁹ <https://www.wikidata.org/wiki/User:Frettiebot>.

the link to NKC ARs into WD items already linked from NKC ARs through the 024 MARC 21 field (the same procedure suggested in chapter 2.1), and also converts some ARs' fields into WD statements (mainly field 374 "Occupation" into P106 "occupation" and field 372 "Field of Activity" into P101 "field of work"). As of now, the NKC case represents the most accomplished case of routinary data synchronisation from an AF to WD.

GND (P227 in WD): GND, currently the second most-linked AF in WD (besides only VIAF), has been tightly connected to the Wikimedia community since before its birth (cf. Wikimedia Deutschland 2025): in 2005 German Wikipedia users started collecting mistake reports for PND, one of the AFs merged together in 2012 to create the GND; since 2016 some of these users have been provided editing access to GND with *Katalogisierungslevel* "Tp4";⁵⁰ after twenty years, German Wikipedia users are still active in connecting their articles to GND (708,640 articles link to GND as of now⁵¹) and in sending monthly lists of mistake reports to the Bavarian State Library (BSB) for people died in 1850 or earlier and to the German National Library (DNB) for people died after 1850 or still alive;⁵² unfortunately, the backlog of reports yet to be checked is of three years for BSB and one year for DNB as of now.⁵³

LCNAF (P244 in WD): since 2023 a bot⁵⁴ has been periodically adding the link to LCNAF ARs into WD items already linked from LCNAF ARs through the 024 and/or 670 MARC 21 fields, and generating an internal report that allows the bot maintainers to find and correct some mistakes in LCNAF ARs themselves (Muratsubaki Campany et al. 2024); the Python code of the bot is publicly available on GitHub.⁵⁵

SBN (P396 in WD): the Central Institute for the Union Catalogue of Italian Libraries and Bibliographic Information (ICCU), the institution coordinating the SBN network, has had a collaboration agreement with Wikimedia Italia (WMIT) since 2015; since the renewal of OPAC SBN's interface in 2021, the joint effort of Wikidata users and ICCU to improve the reconciliation of WD items and SBN AFs has intensified, especially thanks to ICCU's decision to make visible the vast majority of ARs (Pellizzari di San Girolamo 2024c). Since 2023 the collaboration has added to its focuses the massive enrichment of SBN ARs with data from WD so as to bring them from S2 to S0; a first group of ca. 16200 ARs has been enriched on July 29, 2024 (Ravelli 2024); moreover, new reconciliation methods, both manual and semiautomatic, have brought a significant growth of P396 values in WD, from 132,946 at the end of 2023⁵⁶ to 193,035 as of now.⁵⁷ The collaboration is regularly documented in WD.⁵⁸

⁵⁰ <https://de.wikipedia.org/wiki/Hilfe:GND>.

⁵¹ <https://de.wikipedia.org/w/index.php?oldid=258099995>.

⁵² <https://de.wikipedia.org/wiki/Wikipedia:GND/Fehlermeldung>.

⁵³ <https://de.wikipedia.org/w/index.php?oldid=257735082>.

⁵⁴ <https://www.wikidata.org/wiki/User:LccnBot>.

⁵⁵ <https://github.com/lcnetdev/lccn-wikidata-bot>.

⁵⁶ <https://www.wikidata.org/w/index.php?oldid=2038639439&action=edit>.

⁵⁷ <https://www.wikidata.org/w/index.php?oldid=2372669213&action=edit>.

⁵⁸ https://www.wikidata.org/wiki/Wikidata:Gruppo_Wikidata_per_Musei_Archivi_e_Biblioteche/SBN.

Parsifal (P12458 in WD): the AFs of some libraries of the network Unione Romana Biblioteche Ecclesiastiche (URBE) have started being reconciled with WD in 2018; since 2019 a growing number of cataloguers have been directly involved in editing WD, reconciling ARs with WD items through MnM and using the lists of CVs to find and correct mistakes in ARs (Pellizzari di San Girolamo 2024b). Since the creation of the collective catalog of URBE libraries, Parsifal, in 2022, the involvement of librarians in WD has further increased, as shown by a learning day on May 14, 2024⁵⁹ and a meeting of 30+ librarians on July 3, 2025.⁶⁰ The collaboration is regularly documented in WD.⁶¹

5.2 Authority files using Wikibase

Wikibase is the open-source software used by Wikidata;⁶² it consists of a set of MediaWiki extensions that allow to collect structured data in multiple languages with a flexible data model. The possible use of Wikibase as a platform to store an entire AF has long been investigated. In 2019 OCLC published a report regarding a Wikibase prototype used by librarians to create linked data to describe resources (Godby et al. 2019). In 2020 the DNB in collaboration with Wikimedia Deutschland (WMDE) started the development of “a Wikibase-based GND”⁶³ and, in the same year, the French National Library (BNF) and the Agence bibliographique de l’enseignement supérieur (ABES) recommended to use Wikibase as the platform for the development of the Fichier National d’Entités (FNE), a new AF merging the two pre-existing ones, BNF and IDREF, as part of the Transition bibliographique project.⁶⁴ However, the FNE project has been terminated in 2023, judging Wikibase an unsuitable solution, mainly due to the insufficient speed of massive data upload via API,⁶⁵ an issue which has not been overcome yet as of now.⁶⁶ The project of a Wikibase-based GND is still ongoing as of now;⁶⁷ in 2024, however, the final aim of having the 10 M GND entities in two distinct but synchronised platforms, the present one and the Wikibase instance, was perceived as «far away», and «bulk import and export of data» was among the *desiderata*; presently Wikibase is being used mainly to transform GND documentation into structured data (Fischer 2024). OCLC used Wikibase also in the CONTENTdm Pilot Project (Bahnmann et al. 2021), but finally discarded it for the development of the Shared Entity Management Infrastructure, mainly due to the aforementioned low speed of massive data upload (Phillips and

⁵⁹ https://www.wikidata.org/wiki/Wikidata:Gruppo_Wikidata_per_Musei_Archivi_e_Biblioteche/URBE/Formazione.

⁶⁰ <https://en.pusc.it/article/wikilib-bibliotecari-romani-e-vaticani>; <https://www.iccu.sbn.it/it/eventi-novita/novita/LIC-CU-alla-giornata-di-WikiLib/>.

⁶¹ https://www.wikidata.org/wiki/Wikidata:Gruppo_Wikidata_per_Musei_Archivi_e_Biblioteche/Parsifal.

⁶² <https://wikiba.se/>; <https://www.mediawiki.org/wiki/Wikibase>.

⁶³ <https://wiki.dnb.de/spaces/GND/blog/2020/03/07/167019461/Report+GND+meets+Wikibase+2>.

⁶⁴ <https://web.archive.org/web/20200715115414/https://www.transition-bibliographique.fr/wp-content/uploads/2020/07/synthese-preuve-concept-fne.pdf>.

⁶⁵ <https://web.archive.org/web/20231008163559/https://www.transition-bibliographique.fr/2023-09-26-suspension-du-projet-fne-et-perspectives-nouvelles/>.

⁶⁶ <https://phabricator.wikimedia.org/T287164>.

⁶⁷ https://web.archive.org/web/20250210173054/https://gnd.network/Webs/gnd/DE/Projekte/projekte_node.html.

Washington 2022, 46); in 2022 the project has been officially launched⁶⁸ as WorldCat Entities.⁶⁹ Despite this, there are ongoing projects using Wikibase to store authority data.

Czech Republic: a Wikibase instance was created by WMCZ in 2019 with a part of the NKC AF as an intermediate step before importing data into WD (cf. chapter 5.1).⁷⁰ Further use of Wikibase for other NLCZ databases is one of the future possible areas of collaboration between NLCZ and WMCZ (Jansová et al. 2024, 8, 15); e.g. in 2024–5 TDKIV (the Czech Explanatory Terminology Database of Library and Information Science), one of the controlled vocabularies curated by NLCZ, has been converted into a Wikibase instance.⁷¹

Wales: in 2022 the Semantic Name Authority Repository Cymru (SNARC) has been “established in order to provide a central hub for name authority records relating to Wales and in the Welsh language”;⁷² this Wikibase instance has been created and mainly edited by Jason Evans, National Wikimedian at the National Library of Wales (NLW) since 2017,⁷³ and has currently more than 139 thousand items.

Greece: the National Library of Greece (NLG) has worked on a proof of concept using a Wikibase instance as backend to store authority data in RDA/RDF, accessible only to NLG staff, and a public platform⁷⁴ as frontend to visualize data in a user-friendly way and to query them;⁷⁵ the experiment was successful and the project continues to be developed (Zapounidou et al. 2024).

Nigeria: in 2025 the Wikibase instance National Library of Nigeria Semantic Name Authority Repository (NLN SNAR) has been created as part of the Integrating Wikimedia Projects into African Libraries’ Ecosystem project (IWIPALE Project) of African Library and Information Associations and Institutions (AfLIA), supported by the Knowledge Equity Fund of Wikimedia Foundation (WMF) and implemented by the National Library of Nigeria (NLN) and WMDE; it has currently more than 3 thousand items.⁷⁶

6. Conclusions

WD, given its crucial importance in the linked data ecosystem, has become increasingly important for AFs, both for providing PIDs to which ARs can link and for extracting data to enrich ARs; at the same time, WD constantly uses AFs as reliable sources for its data and as authoritative IDs

⁶⁸ <https://www.oclc.org/en/worldcat/entities.html>.

⁶⁹ <https://entities.oclc.org/worldcat/entity>.

⁷⁰ <https://authority.wikimedia.cz/>.

⁷¹ <https://tdkiv.wikibase.cloud/>.

⁷² <https://snarc-llgc.wikibase.cloud/>.

⁷³ <https://www.library.wales/about-nlw/press-and-news/press-releases/2017-press-releases/the-national-library-of-wales-appoint-uks-first-permanent-wikimedian>.

⁷⁴ <https://data.nlg.gr/>.

⁷⁵ <https://data.nlg.gr/query>.

⁷⁶ <https://nlnsnar.wikibase.cloud/>.

for disambiguating its entities (chapter 1). Reconciliation tools, especially MnM and OR, simplify the process of matching ARs and WD items (chapter 2). However, quality issues on both sides can sometimes make it difficult to establish matches: if the identity of the subject of an AR and/or an item is unclear, a match can be established only after it has been clarified (chapter 3). Establishing good procedures to receive mistake reports is crucial for AFs to improve their connection with WD; cataloguers can use WD lists of CVs and self-made queries to find mistakes in both WD items and their ARs; data from WD can be reused in AFs either displaying them through the API or copying them directly into the ARs; vice versa, AFs can also massively import their data into WD, provided that they are released in CC0 license (chapter 4). Relevant collaborations between AFs and WD already exist in Czech Republic, Germany, US, and Italy; the use of Wikibase, WD's software, to store entire AFs has been experimented with mixed results, but presently promising projects exist in Czech Republic, Wales, Greece, and Nigeria (chapter 5). The main challenges in the next years will be involving more cataloguers in WD editing, establishing stable ways of synchronising AFs and WD (including accurate checks on non-biunivocal links) and spreading the use of CC0 license among AFs, so that more ARs can be imported to WD, improving its coverage of the persons described in AFs across the world.

References*

Association of Research Libraries. 2019. 'ARL White Paper on Wikidata: Opportunities and Recommendations'. <https://www.arl.org/wp-content/uploads/2019/04/2019.04.18-ARL-white-paper-on-Wikidata.pdf>.

Bahnemann, Greta, Michael Carroll, Paul Clough, Mario Einaudi, Chatham Ewing, Jeff Mixer, Jason Roy, Holly Tomren, Bruce Washburn, and Elliot Williams. 2021. *Transforming Metadata into Linked Data to Improve Digital Collection Discoverability: A CONTENTdm Pilot Project*. OH: OCLC Research. <https://doi.org/10.25333/fzcv-0851>.

Bargioni, Stefano. 2020. 'From Authority Enrichment to AuthorityBox : Applying RDA in a Koha Environment.' *JLIS.it* 11 (1): 175–89. <https://doi.org/10.4403/jlis.it-12595>.

Bargioni, Stefano, Carlo Bianchini, and Camillo Carlo Pellizzari di San Girolamo. 2023. 'IRIS e Wikidata: Un progetto per una migliore valorizzazione e fruizione degli archivi della ricerca scientifica italiana.' In *Guardando oltre i confini. Partire dalla tradizione per costruire il futuro delle biblioteche. Studi e testimonianze per i 70 anni di Mauro Guerrini*, edited by Giovanni Bergamin and Tiziana Possemato, 31–45. Roma: Associazione italiana biblioteche.

Bartholmei, Stephan, Rachel Franks, James Heilman, Mylee Joseph, Vicki McDonald, Anna Rautnik, Mia Ridge, and Mark Robertson. 2016. 'Opportunities for Academic and Research Libraries and Wikipedia: A Discussion Paper.' <https://www.ifla.org/wp-content/uploads/2019/05/assets/hq/topics/info-society/iflawikipediaopportunitiesforacademicandresearchlibraries.pdf>.

Bianchini, Carlo, Stefano Bargioni, and Camillo Carlo Pellizzari di San Girolamo. 2021. 'Beyond VIAF.' *Information Technology and Libraries* 40 (2). <https://doi.org/10.6017/ital.v40i2.12959>.

Boccone, Alessandra. 2022. 'The Role of the Wikidata Librarian in a Renewed Bibliographical Universe: "Next Generation Metadata", next Generation Librarians.' *JLIS.it* 13 (2): 45–57. <https://doi.org/10.36253/jlis.it-460>.

Candela, Gustavo, Mirjam Cuper, Olga Holownia, Nele Gabriëls, Milena Dobрева, and Mahendra Mahey. 2024. 'A Systematic Review of Wikidata in GLAM Institutions: A Labs Approach.' In *Linking Theory and Practice of Digital Libraries*, edited by Apostolos Antonacopoulos, Annika Hinze, Benjamin Piwowarski, Mickaël Coustaty, Giorgio Maria Di Nunzio, Francesco Gelati, and Nicholas Vanderschantz, 34–50. Cham: Springer. https://doi.org/10.1007/978-3-031-72440-4_4.

Chen, Chen, and Zhong Yuxuan. 2024. 'A Comparative Study on the Approaches of Name Authority Control and Wikidata Identity Management.' In *Knowledge Organization for Resilience in Times of Crisis: Challenges and Opportunities*, edited by Wei Lu, Thiago Henrique Bragato Barros, Lu An, and Daniel Martínez-Ávila, 63–74. Baden-Baden: Ergon. <https://doi.org/10.5771/9783987400476-63>.

Chou, Charlene. 2025. 'Linking Challenges for Pseudonyms: The Studies of VIAF, ISNI, Wikidata, Meridian, and Beyond.' *Cataloging & Classification Quarterly* (2): 1–45. <https://doi.org/10.1080/01639374.2025.2502329>.

* All the links have been last retrieved on July 21, 2025 and rechecked on February 23, 2026; all the references to "now" and "currently" used in the article refer to July 21, 2025.

De Monaco, Sara. 'Il Dizionario degli scrittori italiani contemporanei pseudonimi in Wikidata. Metodologia e risultati'. (Tesi di laurea magistrale, Università di Pisa, 2025), <https://etd.adm.unipi.it/t/etd-05132025-160428>.

De Monaco, Sara, and Camillo Carlo Pellizzari di San Girolamo. 2025. 'Il Dizionario degli scrittori italiani contemporanei pseudonimi in Wikidata. Metodologia e risultati (dati)'. Zenodo. <https://doi.org/10.5281/ZENODO.15474929>.

Dostál, Vojtěch, 'Bringing Czech Authority Files into 21st Century: Integration with Wikidata' (presentation, Data Quality Days, online, September 10, 2021). <https://www.youtube.com/watch?v=JKqZTVisHC4>.

Fischer, Barbara, 'Why We Use Wikibase at the German National Library. What It Takes to Collaboratively Produce Reliable Authority Files' (presentation, Wikidata Days Lisbon, Lisbon, Portugal, November 14, 2024). https://commons.wikimedia.org/wiki/File:Why_we_use_Wikibase_at_the_German_National_Library.pdf.

Godby, Jean, Karen Smith-Yoshimura, Bruce Washburn, Kalan Knudson Davis, Karen Detling, Christine Fernsebner Eslao, Steven Folsom, Xiaoli Li, Marc McGee, Karen Miller, Honor Moody, Craig Thomas, and Holly Tomren. 2019. *Creating Library Linked Data with Wikibase: Lessons Learned from Project Passage*. <https://doi.org/10.25333/faq3-ax08>.

Jansová, Linda, Lenka Maixnerová, and Petra Štátná. 2024. 'Library Data in Wikimedia Projects: Case Study from the Czech Republic.' *O-Bib. Das Offene Bibliotheksjournal / Herausgeber VDB* 11 (4): 1–19. <https://doi.org/10.5282/O-BIB/6081>.

Klein, Maximilian, and Alex Kyrios. 2013. 'VIAFbot and the Integration of Library Data on Wikipedia.' *Code4lib Journal* 22. <https://journal.code4lib.org/articles/8964>.

Muratsubaki Campany, Mary, Kelly Davis, Sasha Frizzell, Elizabeth Plantz, and Crystal Yragui, 'LCCNbot: Creating Connections between the Library of Congress Name Authority File (LCNAF) & Wikidata' (presentation, LD4 Conference, online, October 8, 2024). <https://www.youtube.com/watch?v=T-xKLnOqJ-4>.

Pellizzari di San Girolamo, Camillo Carlo, 'Conflations and Duplications' (presentation., Data Modelling Days, December 1, 2023). https://commons.wikimedia.org/wiki/File:DMD2023_-_Conflations_and_duplications.pdf.

Pellizzari di San Girolamo, Camillo Carlo. 2024a. 'Conflations and Duplications in Wikidata Items: Causes, Detection, Solutions, and Issues.' In *Proceedings of the Wikidata Workshop 2023 Co-Located with 22nd International Semantic Web Conference (ISWC 2023)*. <https://ceur-ws.org/Vol-3640/paper4.pdf>.

Pellizzari di San Girolamo, Camillo Carlo. 2024b. 'Storia della collaborazione tra Wikidata e le biblioteche della Rete URBE nel controllo di autorità.' In *Parsifal. Un modello di collaborazione bibliotecaria per condividere la conoscenza registrata*, edited by Silvano Danieli, 147–62. Firenze: Firenze University Press. <https://doi.org/10.36253/979-12-215-0356-2.17>.

Pellizzari di San Girolamo, Camillo Carlo. 2024c. 'Riconciliare le voci di autorità in SBN con Wikidata. Progressi e prospettive dopo un decennio di lavoro (2013-2023)'. *JLIS.it* 15 (1): 33–44. <https://doi.org/10.36253/jlis.it-573>.

Pellizzari di San Girolamo, Camillo Carlo, 'Data Round-Tripping with Authority Files. An Overview and Some Examples' (presentation, Data Reuse Days, online, February 27, 2025). [https://www.wikidata.org/wiki/File:Data_round-tripping_with_authority_files_\(Data_Reuse_Days_2025\).pdf](https://www.wikidata.org/wiki/File:Data_round-tripping_with_authority_files_(Data_Reuse_Days_2025).pdf).

Phillips, Michael, and Anne Washington. 2022. 'Shared Entity Management Infrastructure di OCLC.' *Biblioteche oggi* 40 (6): 44–51. <https://doi.org/10.3302/0392-8586-202206-044-1>.

Ravelli, Elena, 'Authority file di SBN e Wikidata. Un progetto per l'arricchimento dei dati del Catalogo nazionale' (presentation, Wikidata Days Bologna, Bologna, Italy, November 8, 2024). https://commons.wikimedia.org/wiki/File:Authority_file_di_SBN_e_Wikidata_-_Wikidata_Days_Bologna_2024.pdf.

Sardo, Lucia, and Carlo Bianchini. 2022. 'Wikidata: A New Perspective towards Universal Bibliographic Control.' *JLIS.it* 13 (1): 291–311. <https://doi.org/10.4403/jlis.it-12725>.

Tharani, Karim. 2021. 'Much More than a Mere Technology: A Systematic Review of Wikidata in Libraries.' *The Journal of Academic Librarianship* 47 (2): 102326. <https://doi.org/10.1016/j.acalib.2021.102326>.

Wikimedia Deutschland, 'GND in Wikidata' (presentation, April 14, 2025). https://commons.wikimedia.org/wiki/File:GND_in_Wikidata.pdf.

Zapounidou, Sofia, Lazaros Ioannidis, Michalis Gerolimos, Eftychia Koufakou, and Charalampos Bratsas. 2024. 'Entity Management Using RDA and Wikibase: A Case Study at the National Library of Greece.' *Journal of Library Metadata* 24 (2): 111–31. <https://doi.org/10.1080/19386389.2024.2307208>.

Zhang, Erica, Paromita Biswas, and Iman Dagher. 2023. 'Reflections on the PCC Wikidata Pilot at UCLA Library: Undertaking the PCC Learning Objectives.' *Cataloging & Classification Quarterly* 61 (7–8): 735–72. <https://doi.org/10.1080/01639374.2023.2269416>.

Zhao, Fudie. 2023. 'A Systematic Review of Wikidata in Digital Humanities Projects.' *Digital Scholarship in the Humanities* 38 (2): 852–74. <https://doi.org/10.1093/llc/fqac083>.

Historicising uncertainty in photo archives for semantic modelling purposes

Marilena Daquino^(a)

a) University of Bologna, <https://orcid.org/0000-0002-1113-7550>

Contact: Marilena Daquino, marilena.daquino2@unibo.it

Received: 26 September 2025; **Accepted:** 24 February 2025; **First Published:** 15 May 2026

ABSTRACT

Cultural Heritage (CH) data are inherently ambiguous: attributions shift, authorship is debated, and titles or dates are often uncertain or contested. Cataloguing conventions vary across institutions and time, leading to inconsistencies that affect interoperability. Photo archives amplify these challenges, as early cataloguing practices developed in the absence of shared standards and often treated photographs as secondary to artworks. Linked Open Data (LOD) technologies have opened opportunities to restructure and reconcile such records. However, most models and workflows assume certainty and one-to-one matches between entities, producing overly “clean” interfaces that risk erasing interpretive diversity. Fully reassessing records to resolve uncertainty is rarely feasible, and archivists increasingly face the need to historicise metadata and embrace ambiguity as part of the record itself. This article argues that modelling uncertainty should be treated as a first-class requirement in CH infrastructures, not as a data-cleaning problem. Drawing on cases from PHAROS photo archives, it categorises recurring forms of ambiguity and proposes design principles for representing uncertainty across both data structures and catalogue interfaces, balancing scholarly accuracy with usability.

KEYWORDS

Photo archives; Semantic Web; Uncertainty; Cataloguing.

Introduction

Cultural Heritage (CH) data are intrinsically ambiguous. Attributions shift, authorship is debated, and titles or dates are often uncertain, vague, or questionable. Cataloguing conventions also vary across time and institutions, affecting the accuracy and reliability of information. Photo archives exemplify many of these challenges (Caraffa 2011). Photographic studies consolidated much later than other disciplines, e.g. art historical studies, and only recently has the photograph been recognised as a research object in its own right. As a result, early cataloguing efforts were affected by the absence of standards, terminologies, and shared conventions, which still hinder straightforward cross-institutional interoperability today.

Over the past two decades, Linked Open Data (LOD) technologies, and more recently large language models (LLMs) have offered an unprecedented opportunity to revisit these inherited idiosyncrasies and produce interoperable, shareable data. This process requires dismantling information often flattened or coerced into unsuitable metadata fields, identifying distinct entities and relationships (regardless of their degree of certainty), and reconciling key entities across authorities and institutions to harmonise different traditions, granularities, and perspectives. In fact, most data models and reconciliation workflows assume certainty and one-to-one relationships between entities (Zeng et al. 2021), enabling developers to build clean search interfaces where data appear perfectly interlinked and indexed.

However, reassessing archival records to resolve uncertainties or improve metadata quality is prohibitively costly. For example, photographs attributed simply to “Alinari” are difficult to interpret, as this label could refer to an individual, an atelier, an external photographer hired by the atelier, or other possibilities. This vagueness reflects the shifting priorities of the scholarly community. For a long time, photographers were not considered essential access points to collections, and cataloguers treated their names much like captions. Today, photo archivists must instead historicise their records, seeking a viable compromise that acknowledges uncertainty when stable, factual assertions cannot be made.

This compromise directly affects catalogue user interface design, which must balance data accuracy and usability (Grasso, Daquino, and Renda 2024). Digital infrastructures tend to favour certainty and closure, but this can lead to epistemic loss, e.g. erasing provenance or flattening interpretive diversity. On the other hand, if uncertainty is foregrounded too prominently, it may disorient both scholars and general users (Drucker 2011; Battisti and Daquino 2025). Strategies such as progressive disclosure are therefore needed to introduce complexity gradually and make uncertainty understandable. While the Digital Humanities community has discussed managing uncertainty on a case-by-case basis, there is still no concerted framework for designing catalogue interfaces that are accessible, usable, and meaningfully represent uncertain data.

This article argues that modelling uncertainty must be treated as a first-class requirement in the creation of online catalogues, not as a side effect or a data cleaning problem. Drawing on cases from the photo archives of the PHAROS association, it will categorise common forms of ambiguity in data and propose design principles for representing uncertainty in both data structures and user interfaces.

Related work

Nowadays, catalogues are characterised by some form of authority control, which ensures standardised forms of names, titles, and subjects can be used as coherent access points to collections in organisation and discovery tools (Joudrey et al. 2015). With the widespread adoption of Semantic Web technologies, such standardised names are associated with persistent identifiers like URIs that uniquely identify a real-world entity or concept. As aforementioned, a vague string like “Alinari” may make sense to catalogues’ users as an access point, but it cannot simply be mapped in a one-to-one relation with an entity. Ambiguity and vagueness in cataloguing practices is a recurring topic in the literature (Bates 1986; Brubaker 2002; Olson 2013; Joudrey et al. 2015), whose consequences are amplified and differently handled when multiple catalogues are aggregated in centralised systems. The PHAROS association of art historical archives (Caraffa et al. 2015) is currently working on the realisation of a shared platform, *artresearch.net*, which gathers catalogues of 14 photo archives and harmonises them using Semantic Web technologies. This effort faces a wide range of phenomena characterising archival records, like vagueness, ambiguity, and uncertainty, which have been analysed by data engineers and domain experts in order to define shared policies to handle edge cases and return users with user-friendly views on such complexities.

Scholars and practitioners in the Digital Humanities have also long recognised the importance of preserving ambiguity, uncertainty, and vagueness in Cultural Heritage data—not necessarily to resolve them, but to document and communicate the interpretive debates surrounding them. This has led to a variety of technical approaches spanning knowledge organisation, data reconciliation, and data visualisation. Ontologies such as CIDOC-CRM support the explicit modelling of uncertainty through constructs like *crm:E13_Attribute_Assignment* and qualifiers (Niccolucci and Hermon 2017), while specialised ontologies like the Historical Context (HiCO) Ontology (Daquino and Tomasi 2015) allow attributions and interpretations to be recorded alongside their provenance. On the representation side, techniques such as named graphs (Carroll et al. 2005), conjectural statements (Barabucci, Tomasi, and Vitali 2021), and RDF* (Hartig 2017) enable the encapsulation of contested or provisional assertions without collapsing them into definitive facts. Reconciliation workflows increasingly employ rule-based approaches, graph embeddings, and, more recently, large language models to suggest and qualify matches between certain entities (Zeng et al. 2021). However, reconciliation practices rely on the assumption that a one-to-one match is always possible (if a candidate in the matched authority list is available), which unfortunately clashes with the lack of certainty and the ambiguity that characterises attributionship and cataloguing practices in general.

In data visualisation, uncertainty-aware methods are being explored, particularly in graph-based visualisations, to render confidence levels, alternative hypotheses, and interpretive multiplicity without overwhelming users, thereby maintaining both transparency and usability (Drucker 2011; Battisti and Daquino 2025). Previous studies have also emphasised the importance of linking ontology design practices with user interface design frameworks to improve user experience (Grasso, Daquino, and Renda 2024).

However, while these works acknowledge the presence of uncertainty, they rarely address it as a systematically modelled component carried through from data content to end-user presentation. Strategies such as progressive disclosure (Springer and Whittaker 2019) and generous interfaces

(Whitelaw 2015) have been proposed to mitigate complexity in the interface, but there is still no shared set of practices for representing specific types of uncertainty in a consistent way. In particular, the field lacks empirical studies and efforts to generalise a comprehensive framework that connects the categorisation of ambiguity with corresponding reconciliation and modelling strategies, as well as with suitable visualisation approaches for cultural heritage catalogues.

Classifying ambiguity and uncertainty in photo archives

Catalogues are mirrors of shifting cultural interests and evolving knowledge organisation practices. A wide range of well-known issues affect their quality, becoming especially evident when attempting to reconcile institutional perspectives and traditions through data. The scholarly community has long recognised that archiving and cataloguing are not objective tasks, and each cataloguer inevitably introduces individual choices and idiosyncrasies that contribute further complexity—and sometimes entropy—to the ecosystem of knowledge organisation (Barabucci, Tomasi, and Vitali 2022). These subjective layers complement the inherent uncertainties already present in the knowledge being recorded. While a full inventory of such situations (e.g. homonyms, temporal fuzziness) goes beyond feasible and reasonable classificatory purposes, the classification of situations relevant to (and challenging) semantic data modelling and reconciliation efforts is a pragmatic motivation for the historicisation of photo archives, which need shared policies for expressing, rather than hiding, interpretive diversity in a landscape of digital solutions that favour factuality, flattening, and reticence. In particular, the collection of edge cases presented here exemplifies the uncertainty stemming from the identification of real world entities (people, organisations, places, materials) that affect a straightforward reconciliation between cataloguing data and authority lists.

We can draw a preliminary distinction between uncertainty that stems from gaps in knowledge, uncertainty introduced by cataloguing practices, regardless of how certain or uncertain the underlying knowledge may be at the time of the cataloguing, and uncertainty in knowledge representation with Semantic Web technologies. Specifically, the latter includes mechanisms to cope with the two former situations, which may in turn generate new idiosyncrasies in the way data is processed and interpreted by humans. For the sake of readability, cases are summarised in Table 1 with examples taken from the analysis of PHAROS archival data.¹

¹ Daquino, Mambelli and Kozlov 2026.

Table 1. Classification of uncertainties and ambiguities in photo archival records with examples

Example	Gap in Knowledge	Uncertainty introduced by cataloguers	Uncertainty in knowledge representation
“Brogi/Alinari”	Competing attributions	Reticence Varying reliability	Alternative statements Expressing without asserting
“F.lli Alinari?”	Unknown value	Varying reliability	Assumed certainty Expressing without asserting Reticence in reconciliation Ambiguous anonymous entities
“Alinari”, “Alinari, ?”	Entity ambiguity	Attributional ambiguity	Reconciliation mismatch
“Pseudo Ambrogio di Baldese”	Semantic drift		Reconciliation mismatch Reticence in reconciliation
“Anonymous Florentine, 16 th century”	Semantic drift	Vagueness	Reconciliation mismatch Reticence in reconciliation Assumed certainty
“Auction Christie’s, London 1970”	Unknown value	Coercion	Representation inconsistency Dumping
“Frick collection” vs “Frick Art Reference Library”, “Sistine Chapel” vs “Vatican museums”		Flattening Granularity mismatch Redundancy and divergence	Overlapping categories
“Broomford Manor”	Semantic ambiguity	Flattening	Overlapping categories
“Albumine”			Semantic ambiguity Granularity mismatch

Competing or evolving author attributions, such as “Brogi/Alinari”, are often recorded in catalogues without further assessment. Such a “relativist” approach may reflect cataloguers’ reticence to choose a preferred attribution in the absence of policies, or simply the lack of sufficient evidence. On the one hand, it avoids introducing subjective judgments on top of existing knowledge gaps. On the other hand, it often omits motivations, sources, or explicit acknowledgment of uncertainty, leaving readers without a sense of the reliability of the statements and without a clear explanation of the meaning (users may interpret the string as a collaboration between authors, rather than an alternative attribution). In a Linked Data context, this raises challenges for formally representing (1) alternative hypotheses and (2) the absence of endorsement by the data provider. Formal logic does not allow “either/or” statements, especially when the data do not reflect a true assertion. In practice, cataloguers are reporting hypotheses, possibly narrowed down through analysis or bibliography, but without endorsing or justifying any of them. These cannot therefore be modelled as assertions attributed to the cataloguer but instead require patterns that capture the coexistence of multiple attributions for a given object. However, existing modelling options, such as *crm:E13 Attribute Assignment*, require both an explicit agent responsible for the assignment and a declared value (not an alternative). These solutions do not easily accommodate the expression of hypotheses that fall short of an actual assertion, and do not allow the representation of alternatives. Different

ontology patterns are therefore needed to represent alternatives whose provenance is not directly linked to cataloguers, which act instead as secondary knowledge providers.

A related situation arises when cataloguers record a preferred attribution but mark it as uncertain. For instance, “F.lli Alinari?” signals an identified photographer, but with acknowledged doubt. Here, one may argue the cataloguer both endorses a candidate and encodes an epistemic stance (favoured, speculative, uncertain). In Linked Data this again requires nuance: patterns such as *crm:E13 Attribute Assignment* can be used, but the uncertainty must be captured with custom qualifiers. While this approach is workable, it risks flattening the epistemic subtlety of such cases, since it treats the attribution as a firm assignment rather than a tentative one.

The case of “Alinari” (and “Brogi”) also illustrates attributional ambiguity tied to family names. The string “Alinari” may refer to an individual, an atelier, or both. More problematically, the same string can represent different referents across records, depending on context. In Linked Data, however, URIs must unambiguously denote one entity, and cannot shift meaning depending on context. Reconciliation to external authorities (e.g. VIAF, Wikidata) therefore risks introducing false certainty, forcing ambiguous references into specific entities. Cataloguers may thus refrain from reconciliation, producing duplicate ambiguous entities in authority lists.

Other examples include strings such as “Alinari, ?”, where the family name is recognised, but the individual cannot be identified (e.g. Leopoldo, Giuseppe, or Romualdo). Such labels frame the author as an anonymous member of a family. Reconciliation to external authorities is problematic. One might map the string to a “family group” entity, but this could conflate the intended anonymous group with unrelated individuals. Again, cataloguers may opt for reticence, leading to duplication of uncertain entities in the authority.

Special cases also involve pseudonyms of anonyms. For example, “Pseudo Ambrogio di Baldese” appears as the author of an artwork in several archives. However, scholarship diverges on their identification: some identify the artist as Ventura di Moro (ULAN 500012920), others as Lippo d’Andrea (ULAN 500082343). Aligning the pseudonym with one authority would imply endorsement. When collections are merged, disagreements between institutional perspectives resurface, creating difficulties for reconciliation. In such cases, institutions may avoid external alignment, producing duplicated pseudonymous entities across authority lists.

Similarly, vague attributions such as “Anonymous Florentine, 16th century” define an author in spatio-temporal terms but not as an identifiable person. Institutions may use similar labels to describe entirely different entities. Automatic reconciliation across institutions (since anonymous are poorly represented in authorities) risks conflating them, while refusing reconciliation leads to fragmentation. Policies must therefore choose between reticence and coercion, neither of which fully resolves the ambiguity.

Comparable issues arise in provenance data. Provenance fields typically record an owner (person or institution), a place, and a date range (e.g. “Louvre Museum, 1917–1935”). But real trajectories are often incomplete, and cataloguers must coerce vague or partial evidence into rigid fields. For example, “Auction Christie’s, London, 1970” records not an owner but a sale event where the object was last seen. Such entries disrupt the sequential representation of provenance, complicating reconciliation across institutions, where descriptions may either complement or contradict one another. In practice, cataloguers may resort to dumping complex provenance narratives into text fields, which resist programmatic parsing and semantic modelling.

Granularity mismatches also appear in the identification of entities. Cataloguers sometimes record a collection rather than the holding institution (e.g. “Frick Collection” instead of “Frick Art Reference Library”), treating collections as if they were agents capable of acquisition. This is a form of flattening, where semantically distinct information is collapsed into one field, forcing interpreters to rely on common sense. Similarly, institutions may record either a specific sub-location (“Sistine Chapel”) or a broader institution (“Vatican Museums”), producing overlaps that complicate reconciliation. External authorities often exacerbate the issue. For instance, in Wikidata, the Sistine Chapel is modelled as a building, while the Vatican Museums are an organisation—making it semantically inconsistent to assert a part-whole relation.

Ambiguity also affects entities that blur the line between organisation and place, such as manors (“Broomford Manor”). These can be treated as locations, buildings, or institutions, and often give their names to collections. Here the flattening reflects the inherent semantic ambiguity of the entities themselves.

One last example, this time involving the materiality of described objects, regards a granularity mismatch introduced by semantic technologies. When cataloguers describe a photographic object as “albumine”, one would imagine that the reconciliation to authorities like Getty AAT should be a straightforward operation. Instead, the string may be interpreted (by a human as well as by a program) as the raw material itself (AAT 300411855), the photographic process employing albumine (AAT 300133274), or to the specific product “albumine print” (AAT 300127121). Archives may reconcile to different categories, introducing inconsistencies that, again, require harmonisation.

Towards a model of uncertainty

The survey of ambiguities and uncertainties in cataloguing practices reinforces the broader consensus in the Digital Humanities on the fact that uncertainty is not noise to be eliminated, but a structural feature of cultural heritage data that must be modelled explicitly. Rather than treating it as a side effect of poor cataloguing, uncertainty should be conceptualised as a relational property—linking entities, statements, and interpretations—so that data structures can reflect the interpretive debates and evolving knowledge they embody.

Authority control has historically privileged uniqueness and stability, assuming that entities can and should be reconciled to a single identifier. This bias clashes with humanistic standards, where competing hypotheses, attributions, and evolving interpretations may coexist without being ever resolved. In fact, probabilistic reconciliation approaches, common in data science, suggest one “best” match, but this statistical logic clashes with the epistemic requirements of the humanities, where acknowledging ambiguity is often more faithful to the sources than asserting certainty. Current data structures, however, frequently force premature disambiguation, leading to flattening, coercion, or duplication.

Recent developments in semantic technologies—such as named graphs, the use of `crm:E13 Attribute Assignment` in CIDOC-CRM, and extended formalisms for assertions and conjectures—point to possible ways forward. Instead of collapsing multiple attributions into a single “truth,” catalogues should represent alternative hypotheses explicitly, preserving their provenance and interpretive context.

Some concrete strategies include:

- Minting anonymous or collective entities, e.g. “Alinari family” or “Alinari (collective name),” to avoid false precision and create anchors to gather ambiguous entities under the same definition.
- Using “umbrella terms” to group heterogeneous references under a shared label without assuming identity. E.g. “Alinari” may be an instance of *crm:E33 Linguistic Object* linked to all the aforementioned ambiguous manifestations of people and organisations connected to the common-sense term. This allows one to maintain a relation between all the ambiguous entities via a shareable label, without forcing a precise semantic relation between unknown entities.
- Representing competing attributions as parallel assertions with provenance, ensuring that the origins of each statement remain visible, especially when endorsement is not given. Formalisms for representing competing statements can be found both at RDF level, e.g. using Lists, and with named graphs to separate alternatives.
- Attaching uncertainty scores or qualifiers to mark degrees of confidence or epistemic stance (favoured, tentative, speculative).
- Finally, maintaining reticence, e.g. deliberately withholding reconciliation when ambiguity is irreducible, rather than introducing false certainty.

User interfaces as sites of negotiation

The user interface is the place where ambiguity becomes visible and semantic modelling decisions shape the user experience. The challenge is to render complexity without overwhelming users, especially in the three typical main points of interaction of archives catalogues: facets, lists of results, and individual records. In many cultural heritage platforms, the research experience is reduced to a single input field (e.g. Google-like search input), rather than advanced search options, which shifts the burden of interpretation to users. This design choice requires careful strategies to make ambiguity visible while still supporting intuitive discovery.

Progressive disclosure offers one solution. Canonical or most stable names (e.g. “umbrella terms”) can be displayed by default, with alternatives, variants, and confidence levels accessible on demand, e.g. in hierarchical views or dropdowns. This allows users to navigate familiar, uncluttered interfaces while retaining access to deeper layers of interpretive nuance on demand.

Faceted navigation can also play a critical role. Uncertain or ambiguous entities, such as anonymous groups or umbrella terms, should be included in filters, but visually distinguished (for example, through typographic cues or icons) so that users can recognise their provisional nature. In this way, uncertainty becomes a navigable dimension rather than a barrier.

Finally, record-level views must reconcile clarity with transparency. Users should be able to see both the preferred form of an attribution and the alternative hypotheses, along with provenance details that justify it (e.g. logos of data providers, original non-normalised strings appearing in catalogues). Here too, the key is incremental complexity, exposing certainties and uncertainties without disorienting the reader. To do so, data modelling and interface design must be co-developed. The way uncertainty is modelled at the data level directly constrains how it can be represented in the interface, while interface requirements should inform modelling choices to ensure usability.

Discussion

What emerges from the survey is the need for a framework that balances interoperability with epistemic integrity. Cultural heritage infrastructures should not only connect data across institutions but also safeguard the ambiguities that are intrinsic to the historical record. A model of uncertainty must therefore accommodate multiple perspectives, preserve interpretive openness, and still deliver usable, navigable interfaces to end users. Only by embracing uncertainty as a first-class feature of knowledge representation can digital catalogues both support interoperability and remain faithful to the humanistic traditions they serve.

This perspective also carries implications for the long-term sustainability of cultural heritage systems. Catalogues are never definitive, since new evidence, changing scholarly consensus, and evolving descriptive practices will continue to reshape attributions, identifications, and provenance. Systems designed with rigidity or premature closure risk becoming obsolete as soon as new knowledge emerges. Conversely, infrastructures that can accommodate revision, store alternative hypotheses, and represent degrees of confidence are more resilient. They allow catalogues to evolve while maintaining provenance and interpretive transparency, ensuring that future work builds on, rather than overwrites, existing records.

Finally, modelling uncertainty situates cultural heritage practice within the wider debate on data epistemology in the humanities. Scholars have long argued that digital representations of culture must reflect the partial, contested, and situated character of knowledge rather than aspiring to an impossible neutrality. Approaches that treat uncertainty as a first-class feature align with critical traditions in archival science, historiography, and the digital humanities, which view data not as fixed facts but as interpretations embedded in social, institutional, and technological contexts. In this light, uncertainty is not a flaw but a sign of intellectual honesty, a recognition that cultural heritage is as about what we know as about what we do not know.

Conclusion

Ambiguity and uncertainty are not accidents of cataloguing but inherent features of cultural heritage data. Far from diminishing value, they enrich the record by preserving traces of scholarly debate, shifting conventions, and incomplete knowledge. Treating them as assets rather than errors requires moving beyond data-cleaning logics toward explicit modelling strategies that can capture hypotheses, alternatives, and degrees of confidence.

This study has shown that such an approach is not only possible but necessary. Infrastructures, standards, and policies must be developed with uncertainty at their core, ensuring that digital catalogues remain transparent, revisable, and faithful to the interpretive nature of the humanities. By embedding ambiguity into both data models and user interfaces, we can design systems that are not only interoperable and sustainable, but also intellectually honest.

References

- Barabucci, Gioele, Francesca Tomasi, and Fabio Vitali. 2021. "Supporting complexity and conjectures in cultural heritage descriptions." In *CEUR Workshop Proceedings* 2810: 104-115. <https://hdl.handle.net/11250/2736994>.
- Barabucci, Gioele, Tomasi, Francesca and Fabio Vitali. 2022. "Modeling Data Complexity in Public History and Cultural Heritage." In Noiret, Serge, Mark Tebeau, and Gerben Zaagsma, *Handbook of Digital Public History*. Berlin, Boston: De Gruyter Oldenbourg, 459-74. <https://doi.org/10.1515/9783110430295-041>.
- Bates, Marcia J. 1986. "Subject access in online catalogs: A design model." *Journal of the American society for information science* 37 (6): 357-76. [https://doi.org/10.1002/\(SICI\)1097-4571\(198611\)37:6<357::AID-ASI1>3.0.CO;2-H](https://doi.org/10.1002/(SICI)1097-4571(198611)37:6<357::AID-ASI1>3.0.CO;2-H).
- Battisti, Tommaso, and Marilena Daquino. 2025. "Uncertainty, narrativity, and critical approaches in Digital Humanities information visualisation projects." In *Proceedings of the 21st Conference on Information and Research science Connecting to Digital and Library science*, 1-10. <https://ceur-ws.org/Vol-3937/short1.pdf>.
- Brubaker, Jana. 2002. "Ambiguous authorship and uncertain authenticity: A cataloger's dilemma." *Cataloging & Classification Quarterly* 34 (4): 19-30. https://doi.org/10.1300/J104v34n04_04.
- Caraffa, Costanza, ed. 2011. *Photo archives and the photographic memory of art history*. Berlin: Deutscher Kunstverlag.
- Caraffa, Costanza, Emily Pugh, Tracy Stuber, and Louisa Wood Ruby. 2020. "Pharos: A digital research space for photo archives." *Art Libraries Journal* 45 (1): 2-11. <https://doi.org/10.1017/alj.2019.34>.
- Carroll, Jeremy J., Christian Bizer, Pat Hayes, and Patrick Stickler. 2005. "Named graphs." *Journal of Web Semantics* 3(4): 247-67. <https://doi.org/10.1016/j.websem.2005.09.001>.
- Daquino, Marilena, Mambelli, Francesca, & Kozlov, Artem. 2026. "Many-to-many. Usability challenges of entity reconciliation in art history and photographic studies." *Journal of Documentation* 82 (2): 362–380. <https://doi.org/10.1108/JD-09-2025-0284>
- Daquino, Marilena, and Francesca Tomasi. 2015. "Historical Context Ontology (HiCO): a conceptual model for describing context information of cultural heritage objects." In *Metadata and Semantics Research. MTSR 2015. Communications in Computer and Information Science*, edited by Emmanouel Garoufallou, Richard J. Hartley, and Panorea Gaitanou, 424-36. Cham: Springer. https://doi.org/10.1007/978-3-319-24129-6_37.
- Drucker, Johanna. 2011. "Humanities approaches to graphical display." *Digital Humanities Quarterly* 5 (1): 1-21. <https://dhq.digitalhumanities.org/vol/5/1/000091/000091.html>.
- Grasso, Marco, Marilena Daquino, and Giulia Renda. 2024. "From Ontology Design to User Experience. A methodology to design interfaces for information seeking purposes." *Umanistica Digitale* 18: 53-85. <http://doi.org/10.6092/issn.2532-8816/19038>.

Hartig, Olaf. 2017. "Foundations of RDF* and SPARQL* (An Alternative Approach to Statement-Level Metadata in RDF)." In *Proceedings of the 11th Alberto Mendelzon International Workshop on Foundations of Data Management and the Web. CEUR Workshop Proceedings 1912*, edited by Juan Reutter and Divesh Srivastava. <https://ceur-ws.org/Vol-1912/>.

Joudrey, Daniel N., Arlene G. Taylor, and David P. Miller. 2015. *Introduction to cataloging and classification*. London: Bloomsbury Publishing.

Nicolucci, Franco, and Sorin Hermon. 2017. "Expressing reliability with CIDOC CRM." *International Journal on Digital Libraries* 18 (4): 281-87. <https://doi.org/10.1007/s00799-016-0195-1>.

Olson, Hope A. 2013. *The power to name: locating the limits of subject representation in libraries*. Berlin: Springer Science & Business Media.

Springer, Aaron, and Steve Whittaker. 2019. "Progressive disclosure: empirically motivated approaches to designing effective transparency." In *Proceedings of the 24th international conference on intelligent user interfaces*, 107-20. <https://doi.org/10.1145/3301275.3302322>.

Whitelaw, Mitchell. 2015. "Generous interfaces for digital cultural collections." *Digital Humanities Quarterly* 9 (1). <https://dhq.digitalhumanities.org/vol/9/1/000205/000205.html>.

Zeng, Kaisheng, Chengjiang Li, Lei Hou, Juanzi Li, and Ling Feng. 2021. "A comprehensive survey of entity alignment for knowledge graphs." *AI Open* 2: 1-13. <https://doi.org/10.1016/j.aiopen.2021.02.002>.

Self-Publishing and libraries: cataloguing challenges and legal deposit issues in self-published books

Elda Merenda^(a)

a) National central library of Rome, <https://orcid.org/0000-0002-5727-8690>

Contact: Elda Merenda, elda.merenda@cultura.gov.it

Received: 30 December 2025; **Accepted:** 16 March 2026; **First Published:** 15 May 2026

ABSTRACT

This paper analyzes the phenomenon of self-publishing within the context of contemporary publishing, with particular focus on its impact on the library system, cataloguing practices, and legal deposit policies. Starting with the quantitative and structural development of self-publishing in Italy, the study examines the critical issues that arise in relation to bibliographic identification, metadata management, and the definition of publishing responsibilities, highlighting the difficulties libraries face in handling materials lacking traditional editorial mediation. The essay analyzes the role of the ISBN, the limitations of authority systems, and the issues related to the availability and reliability of information on authors, as well as the impact of these factors on cataloging and preservation practices. Particular attention is devoted to the legal framework of legal deposit and its implementation challenges in the case of self-published works, especially regarding digital content. Overall, the article offers a reflection on self-publishing as a structural phenomenon of the contemporary publishing landscape and as a test case for libraries' ability to adapt tools, criteria, and functions to the transformation of cultural production models.

KEYWORDS

Self publishing; Legal deposit; Authority control; Publisher; Bibliographic metadata.

Self publishing e biblioteche: criticità catalografiche e problemi di deposito legale dei libri autopubblicati

ABSTRACT

Il contributo analizza il fenomeno del self publishing nel contesto dell'editoria contemporanea, con particolare attenzione alle ricadute che esso produce sul sistema bibliotecario, sulle pratiche catalografiche e sulle politiche di deposito legale. A partire dall'evoluzione quantitativa e strutturale dell'autopubblicazione in Italia, il saggio esamina le criticità che emergono in relazione all'identificazione bibliografica, alla gestione dei metadati e alla definizione delle responsabilità editoriali, evidenziando le difficoltà incontrate dalle biblioteche nel trattamento di materiali privi di mediazione editoriale tradizionale. Vengono analizzati il ruolo dell'ISBN, i limiti dei sistemi di autorità e le problematiche connesse alla reperibilità e all'affidabilità delle informazioni sugli autori, nonché l'impatto di tali fattori sulle pratiche di catalogazione e conservazione. Particolare attenzione è dedicata al quadro normativo del deposito legale e alle sue criticità applicative nel caso delle pubblicazioni autoprodotte, soprattutto in relazione ai contenuti digitali. Nel complesso, l'articolo propone una riflessione sul self publishing come fenomeno strutturale del panorama editoriale contemporaneo e come banco di prova per la capacità delle biblioteche di adattare strumenti, criteri e funzioni alla trasformazione dei modelli di produzione culturale.

PAROLE CHIAVE

Autopubblicazione; Deposito legale; Controllo di autorità; Editore; Controllo bibliografico universale.

Il self publishing nel panorama editoriale contemporaneo

Negli ultimi venti anni il self publishing ha assunto un ruolo sempre più rilevante nel panorama editoriale italiano, contribuendo in modo significativo alla ridefinizione delle dinamiche della produzione libraria. Il fenomeno, inizialmente marginale, ha conosciuto una rapida espansione a livello globale grazie allo sviluppo delle tecnologie digitali, alla riduzione dei costi di stampa e distribuzione e alla crescente accessibilità degli strumenti di pubblicazione, fino a configurarsi come uno dei settori più dinamici – e al tempo stesso più problematici – dell’editoria contemporanea.

Con il termine self publishing si indica l’insieme delle pratiche attraverso le quali un autore pubblica autonomamente la propria opera, assumendo la responsabilità dell’intero processo editoriale: dalla scrittura alla revisione, dall’impaginazione alla realizzazione grafica, fino alla promozione e alla distribuzione. A differenza del modello editoriale tradizionale, fondato sulla selezione preventiva dei testi in coerenza con una precisa linea editoriale e sull’investimento economico dell’editore che svolge un preciso ruolo di mediazione culturale (Peresson 2016), l’autopubblicazione trasferisce sull’autore tanto il controllo creativo su tutti i passaggi quanto il rischio economico, configurando un modello produttivo profondamente diverso.¹

Questa tendenza,² diffusa negli Stati Uniti d’America già a partire dai primi anni Duemila, si è progressivamente affermata anche in Europa, fino a rappresentare una quota rilevante del mercato editoriale globale. In Italia il fenomeno ha iniziato a consolidarsi a partire dal 2008, quando l’*E-spresso*, in collaborazione con il gruppo Feltrinelli e la Scuola Holden, ha lanciato in via sperimentale il sito *ilmiolibro.it*, che ancora è tra le piattaforme che offrono servizi editoriali a pagamento più utilizzate. A questa esperienza hanno fatto seguito numerose altre iniziative, come *narcissus.me* e *youcanprint.it*, oltre ai colossi internazionali, tra cui Amazon Kindle Direct Publishing e *lulu.com*. Tali piattaforme offrono un insieme articolato di servizi che vanno ben oltre la semplice pubblicazione, includendo anche consulenze editoriali, servizi di grafica, strumenti di promozione e distribuzione multicanale, configurandosi come intermediari tecnologici più che come editori in senso tradizionale (Garzia 2023, 49).³

I dati più recenti confermano la rilevanza quantitativa del fenomeno e la sua crescita costante. Secondo il *Rapporto sullo stato dell’editoria in Italia* dell’Associazione Italiana Editori, nel 2023 sono stati pubblicati in regime di self publishing 12.968 titoli a stampa, pari a circa il 16% della produzione complessiva, ai quali si aggiungono 11.862 titoli in formato digitale (AIE 2024, 21). Dal

¹ Il ricorso all’autopubblicazione attraverso forme diverse di finanziamento (sottoscrizioni degli amici e filantropi, pagamento di tasca propria) è avvenuto in passato anche per alcuni autori italiani e stranieri di fama internazionale. Si pensi che il primo volume dell’opera *À la recherche du temps perdu* di Marcel Proust venne rifiutato da ben 3 editori, prima che l’autore decidesse di autopubblicare il dattiloscritto pagando di tasca propria le spese di pubblicazione, stampa e pubblicità. Tra gli altri possiamo ricordare anche Italo Svevo e Carlo Emilio Gadda in Italia, o ancora Emily Dickinson, Edgar Allan Poe, Mark Twain e Walt Whitman (Garzia 2023, 49).

² L’autore diventa auto-editore e pur avvalendosi molto spesso della collaborazione di altri professionisti del settore mantiene l’ultima parola su tutti i passaggi del processo produttivo (Zita 2016).

³ Per un confronto sul funzionamento e i servizi offerti dalle piattaforme in uso in Italia vedi (Moretti 2013; Garzia 2023, 53-8). Negli ultimi anni invece hanno iniziato a diffondersi software di scrittura assistita che grazie all’impiego dell’Intelligenza Artificiale guidano l’utente in tutto il percorso di costruzione del libro, dall’impostazione del progetto all’impaginazione. Il primo prodotto italiano di questo tipo si chiama PlottyBot ed è nato dall’esperienza di un gruppo di selfpublisher (<https://plottybot.com/>).

punto di vista dei contenuti, nei libri autopubblicati prevale la narrativa, soprattutto nel formato digitale, mentre nel cartaceo si registra una maggiore diversificazione, con una crescita significativa della saggistica pratica e dei libri per l'infanzia. Se si considera anche il valore economico generato attraverso i tradizionali canali di vendita del libro, ossia le librerie fisiche e online inserite nei circuiti distributivi ordinari, il comparto del self publishing può essere assimilato a un editore di medio-grandi dimensioni. Tuttavia, la sua effettiva consistenza rimane difficile da quantificare, sia per la presenza di una quota significativa di produzione non tracciata – in particolare nel caso delle piattaforme che non utilizzano l'ISBN⁴ – sia per l'eterogeneità dei modelli distributivi adottati. Parallelamente, il self publishing è divenuto un importante bacino di osservazione per l'editoria tradizionale, che sempre più spesso individua in questo ambito nuovi autori e tendenze narrative da valorizzare attraverso percorsi di pubblicazione convenzionali.⁵ Nonostante la sua diffusione, il self publishing rimane un ambito di difficile definizione dal punto di vista bibliografico e catalografico. L'eterogeneità dei modelli produttivi, l'assenza di filtri editoriali omogenei e la presenza di circuiti distributivi non sempre tracciabili pongono problemi significativi in termini di descrizione, conservazione e accesso. In questo senso, il fenomeno si colloca pienamente all'interno di quella che è stata definita la “quarta generazione dell'editoria”, caratterizzata da processi di disintermediazione, dalla centralità delle piattaforme digitali e da una ridefinizione profonda dei ruoli tradizionali di autore, editore e lettore (Ragone 2005; 2023, 8).

L'impatto del self publishing sul sistema bibliotecario

Il fenomeno del self publishing non può essere analizzato esclusivamente sotto il profilo economico o produttivo, ma richiede di essere osservato anche in relazione all'impatto che esso esercita sul sistema bibliotecario. Fin dalle sue prime manifestazioni, infatti, l'autopubblicazione ha posto alle biblioteche una serie di nuove questioni, costringendo i professionisti dell'informazione a confrontarsi con tipologie documentarie caratterizzate da elementi di forte discontinuità rispetto alla produzione editoriale tradizionale.

Uno degli aspetti maggiormente problematici riguarda la qualità bibliografica di queste pubblicazioni. Sebbene negli ultimi anni si registri un progressivo miglioramento sotto il profilo formale e contenutistico, molti titoli autopubblicati continuano a presentare carenze significative sul piano

⁴ È il caso, ad esempio, di Amazon Kindle Direct Publishing che consente la pubblicazione di ebook senza codice ISBN, ma assegnando all'opera un ASIN (Amazon Standard Identification Number) che è esclusivo della sua piattaforma. Sempre Amazon KDP consente la stampa *on demand* assegnando un ISBN proprietario gratuito, di cui Amazon è titolare e che è vincolato contrattualmente alla sua piattaforma.

⁵ Si pensi alla trilogia erotica di *50 sfumature* dell'autrice britannica Erika Leonard James, prima pubblicata sul blog della scrittrice e distribuita in *print on demand*, poi divenuta un vero e proprio caso mondiale. Anche in Italia non sono mancati i casi di successo, come quelli di Federico Moccia che già nel 1992 non trovando una casa editrice interessata a *Tre metri sopra il cielo* decise di pubblicare il romanzo a sue spese, stampandone poche copie che distribuì gratuitamente tra gli studenti dei licei romani. Più di recente Anna Premoli e Martin Rua sono diventati entrambi autori per Newton Compton, o ancora Amabile Giusti, che pubblica sia con Mondadori sia con Amazon (Garzia 2023, 60-1). Non mancano invece gli esempi di self publisher di successo che hanno deciso di continuare a pubblicare le loro opere come autori indipendenti: è il caso di Rita Carla Francesca Monticelli, autrice di fantascienza che ha venduto migliaia di copie in italiano e decine di migliaia in inglese (Zita 2016).

redazionale e tipografico, con testi non sottoposti a revisione professionale e privi di un apparato editoriale strutturato, dal momento che la qualità del prodotto finale varia molto a seconda della piattaforma con cui si sceglie di pubblicare. Ciò ha contribuito a consolidare, almeno in una fase iniziale, una diffusa diffidenza da parte delle biblioteche, che spesso hanno percepito il self publishing come un settore marginale o qualitativamente debole rispetto alla produzione editoriale tradizionale (O'Dair 2018, 146; Dinotola 2021, 401).

Tuttavia, la progressiva crescita del fenomeno e la sua rilevanza culturale hanno reso sempre meno sostenibile un atteggiamento di esclusione aprioristica. In particolare, la crescente visibilità di alcuni autori autopubblicati e il ruolo sempre più centrale delle piattaforme digitali come luoghi di scoperta e di mediazione culturale hanno indotto molte biblioteche a riconsiderare le proprie politiche di selezione, spostando progressivamente l'attenzione verso nuovi indicatori di rilevanza, quali il successo presso il pubblico, le recensioni online o l'impatto sui social networks (Pacer 2013, 33). Tali indicatori tendono oggi ad affiancarsi ai criteri tradizionali di selezione bibliografica, fondati principalmente su requisiti di autorevolezza, accuratezza, attualità e rilevanza scientifica, sul prestigio della casa editrice e sulla qualità formale ed editoriale della pubblicazione, sulla disponibilità di recensioni su riviste specializzate e di citazioni in opere altrettanto autorevoli o repertori bibliografici, nazionali ed internazionali. Tali caratteristiche risultano tuttavia frequentemente assenti o fortemente carenti nelle opere autopubblicate, a tal punto da costituire uno dei principali fattori che storicamente ne hanno limitato l'ingresso nei patrimoni bibliotecari.

Allo stesso tempo, il self publishing ha posto questioni rilevanti anche sul piano della gestione e della conservazione delle collezioni, ponendo nuove e rilevanti sfide per il controllo bibliografico universale. In ragione del carattere eterogeneo e spesso instabile di queste produzioni, le risorse autopubblicate figurano tra quelle più frequentemente prese in considerazione nell'ambito delle politiche di selezione e scarto, soprattutto in un contesto caratterizzato da pesanti limitazioni di spazio e di risorse e dalla cronica carenza di personale adeguatamente formato. Non a caso, essi rientrano spesso tra le tipologie bibliografiche per le quali si valuta l'adozione di strategie di conservazione selettiva o di campionamento (AIB. Commissione nazionale biblioteche e servizi nazionali 2019, 429).

Un elemento critico è costituito dall'esclusione dei libri autopubblicati dai canali distributivi tradizionalmente utilizzati dalle biblioteche per le acquisizioni: essi non figurano sistematicamente nei cataloghi dei principali distributori a cui le biblioteche italiane fanno ricorso per i loro acquisti, come Messaggerie Libri, ALI – Agenzia Libreria Internazionale o Casalini Libri, i quali operano secondo modelli distributivi consolidati, basati su criteri di selezione editoriale, e che garantiscono normalmente la disponibilità di metadati strutturati, informazioni bibliografiche complete e strumenti di selezione affidabili. In assenza di tali strumenti, il processo di acquisizione risulta più oneroso e richiede un significativo investimento di tempo e competenze da parte del personale bibliotecario, che deve effettuare la selezione direttamente sulle piattaforme online. Il modello editoriale tradizionale, infatti, fornisce alle biblioteche una serie di servizi fondamentali – dalla produzione dei metadati alla mediazione critica attraverso recensioni professionali – che nel caso del self publishing sono spesso assenti o fortemente disomogenei. Nel contesto delle biblioteche statunitensi, inoltre, uno dei principali ostacoli all'inclusione sistematica di tali pubblicazioni nelle collezioni risiede proprio nella mancanza della scheda CIP (Cataloging in Publication), servizio che, sin dal 1971, costituisce uno strumento essenziale per la catalogazione preventiva delle opere

edite e che risulta generalmente inaccessibile alle pubblicazioni in regime di self publishing, in quanto non riconducibili a un editore formalmente accreditato.⁶ Ne consegue la necessità, per le biblioteche, di procedere con maggiore frequenza a una catalogazione ex novo delle opere autopubblicate, con un aggravio significativo sul piano organizzativo e gestionale. Tuttavia, già nell'ultimo decennio si sta assistendo a una inversione di tendenza, soprattutto in numerose biblioteche pubbliche che hanno progressivamente iniziato ad accogliere nelle proprie collezioni, grazie alla diffusione di recensioni online e di nuovi sistemi di valutazione partecipata, opere autopubblicate, in particolare di autori locali,⁷ contribuendo alla condivisione delle relative notizie bibliografiche attraverso l'Online Computer Library Service (OCLC). Questo ha consentito non solo di rispondere alla domanda crescente da parte dell'utenza, ma nello stesso tempo di garantire quel pluralismo documentario che rischia di essere compromesso dalla esclusione arbitraria della produzione autopubblicata, che comunque rappresenta un segmento significativo della produzione editoriale contemporanea. Parallelamente molte biblioteche pubbliche statunitensi hanno sviluppato servizi mirati specificamente rivolti agli aspiranti autori, che includono corsi di scrittura creativa, workshop sull'editing, consulenze su ISBN e copyright, assistenza editoriale all'impaginazione e piattaforme di pubblicazione digitale integrate nei cataloghi bibliotecari (Holley 2014, 5). La Los Angeles Public Library, ad esempio, promuove il programma *LAPL Writes*, che mette a disposizione risorse per la scrittura, strumenti di produzione editoriale e la possibilità di distribuire ebook autopubblicati attraverso l'*Indie Author Project*.⁸ Analogamente, la New York Public Library organizza workshop e iniziative informative dedicate alla pubblicazione e al self publishing, come il ciclo di eventi *Pathways to Publishing*.⁹

Dalla selezione alla descrizione: problemi di identificazione, catalogazione e responsabilità editoriale

La presenza dei libri autopubblicati nelle collezioni delle biblioteche pone il bibliotecario di fronte alla difficoltà di trattare correttamente sul piano catalografico risorse spesso prive di metadati e di informazioni bibliografiche chiare, che costringono a mettere in discussione alcuni presupposti strutturali su cui si fondano i modelli tradizionali di controllo bibliografico, in particolare la chiara identificazione dei ruoli editoriali e la stabilità delle informazioni descrittive.

Codici di catalogazione quali le *Regole italiane di catalogazione* (Reicat) e le *Anglo-American Cataloguing Rules – 2nd Edition* (AACR2) individuano nel frontespizio la fonte primaria di informazione, alla quale il catalogatore deve fare riferimento in via prioritaria per la trascrizione degli elementi essenziali della descrizione, ricorrendo ad altre fonti interne o esterne solo in sua assenza o

⁶ La catalogazione partecipata in SBN non si avvale di un servizio CIP offerto da un programma nazionale paragonabile al servizio di OCLC (Manzoni 2023, 65-6). Per questa ragione la mancanza sistematica di una scheda CIP non costituisce il principale ostacolo all'acquisizione di questi titoli all'interno del contesto italiano, dove la biblioteca che cataloga per prima il libro si assume l'onere di farlo per tutte le altre biblioteche della comunità.

⁷ Molto utilizzati in ambito statunitense risultano essere Goodreads (<https://www.goodreads.com/about/us>) e LibraryThing (<https://www.librarything.com/>) (Puglisi 2020, 608).

⁸ https://www.lapl.org/books-edia/lapl-writes?utm_source=chatgpt.com.

⁹ https://www.showclix.com/event/pathways?utm_source=chatgpt.com e <https://www.nypl.org/node/5660>.

in caso di dati incompleti.¹⁰ Analogamente, nello standard *Resource Description and Access* (RDA) il frontespizio è qualificato come fonte preferita di informazione per le pubblicazioni monografiche a stampa, pur all'interno di un impianto più flessibile che riconosce la possibilità di utilizzare qualsiasi fonte ritenuta affidabile, purché adeguatamente giustificata e documentata mediante note.¹¹ La mancanza di frontespizio, tuttavia, costituisce una caratteristica ricorrente delle opere autopubblicate e questo costringe il catalogatore a ricavare le informazioni bibliografiche da fonti complementari, quali la copertina, i preliminari e il colophon, o se necessario da fonti esterne, richiedendo un impegno maggiore rispetto alla catalogazione dei prodotti dell'editoria tradizionale. Un ulteriore elemento di criticità risiede nella sovrapposizione tra le funzioni di autore, editore e distributore, coincidenti nella stessa figura, e la mancanza di indicazioni rispetto al luogo e alla data di pubblicazione: i dati che si evincono dalla pubblicazione sono per lo più quelli di copyright legati al nome dell'autore. In tali circostanze, l'attività catalografica richiede un livello più elevato di analisi e di valutazione critica delle fonti disponibili, al fine di identificare correttamente gli elementi descrittivi e di restituirli in forma coerente e trasparente, anche mediante il ricorso a note esplicative che documentino le scelte adottate (Tuncer e Reed 2019, 18-20).

Un altro aspetto problematico riguarda una accurata compilazione dei dati di autorità: nella maggior parte dei casi, infatti, si tratta di autori per i quali è molto difficile reperire informazioni utili a compilare notizie biografiche esaustive e a gestire eventuali casi di omonimia, poiché, a differenza di quello che accade per l'editoria tradizionale, raramente i libri autopubblicati riportano in quarta di copertina una biografia sintetica dell'autore. E questo costringe il catalogatore a fare spesso ricorso a fonti diverse da quelle abitualmente consultate, cartacee e non, trovando più facilmente informazioni attraverso social networks o blog personali. Inoltre, gli autori operano frequentemente sotto pseudonimo, utilizzano varianti onomastiche non normalizzate o forniscono dati autoreferenziali non verificabili, rendendo complessa l'attribuzione univoca delle opere e la costruzione di punti di accesso controllati. Questa criticità assume particolare rilievo anche in relazione ai principali archivi di autorità condivisi, nazionali e internazionali, a cominciare da VIAF e ISNI, che si fondano sulla stabilità e verificabilità delle informazioni biografiche. L'assenza di dati affidabili limita l'inclusione degli autori autopubblicati nei repertori di autorità e nei sistemi di identificazione persistente, contribuendo a una marginalizzazione informativa che non riflette necessariamente la rilevanza culturale o quantitativa della loro produzione e a una progressiva frammentazione del patrimonio informativo.

Il tema dell'ISBN riveste un ruolo centrale nell'analisi delle criticità connesse al self-publishing, in quanto si colloca in una posizione di raccordo tra identificazione bibliografica, responsabilità editoriale e pratiche di catalogazione. In una prima fase l'ISBN ha contribuito all'identificazione dei titoli autopubblicati, dal momento che l'agenzia ISBN per l'area di lingua italiana ha riservato, già dal 2015, il prefisso 979-12 per gli authorpublishers e i richiedenti occasionali (Puglisi 2020, 609).

¹⁰ Per le fonti delle informazioni prescritte nel contesto catalografico italiano si veda Reicat, capitolo 3 (https://norme.iccu.sbn.it/index.php?title=Reicat/Parte_I/Capitolo_3) e la Guida alla catalogazione in SBN – Materiale moderno, capitolo 0E (https://norme.iccu.sbn.it/index.php?title=Guida_moderno/Descrizione/Capitolo_generale/Fonti_d%27informazione). Per le AACR2: Anglo-American Cataloguing Rules, Second Edition. 2002 revision. Chicago: American Library Association; Ottawa: Canadian Library Association; London: Chartered Institute of Library and Information Professionals, rule 2.0B1.

¹¹ RDA: *Resource Description and Access*, 2.2.2.2.

Tuttavia, a partire dal 2021 il prefisso ISBN 979-12 è entrato ufficialmente in vigore anche per gli editori tradizionali per il progressivo esaurimento delle combinazioni disponibili all'interno del prefisso storico 978-88, assegnato all'area linguistica italiana fin dall'introduzione dello standard ISBN a 13 cifre. Per questa ragione il prefisso 979-12 non è più distintivo di una particolare categoria di pubblicazioni o di soggetti editoriali, come è stato all'inizio della sua introduzione, poiché adesso accomuna editori tradizionali, microeditori e operatori del self publishing.¹²

Nella tradizione biblioteconomica, l'ISBN non rappresenta soltanto un codice identificativo, ma costituisce uno degli elementi attraverso cui distinguere in modo univoco ciascuna edizione di un'opera e collegarla al soggetto responsabile della sua pubblicazione.

Nel contesto del self publishing, tale presupposto risulta frequentemente indebolito. L'adozione di ISBN forniti da piattaforme di autopubblicazione o da aggregatori commerciali (come nel caso di Amazon) introduce una frattura tra responsabilità intellettuale e responsabilità editoriale: l'autore rimane il principale responsabile del contenuto, ma l'ISBN risulta intestato a un soggetto terzo che non esercita un reale ruolo editoriale. In tali contesti, l'ISBN può assolvere una funzione puramente tecnica di identificazione del prodotto all'interno di un ecosistema chiuso, senza corrispondere a una reale assunzione di responsabilità editoriale.

Autopubblicazione e deposito legale: criticità normative e applicative

Un numero rilevante di libri autopubblicati arriva nelle biblioteche italiane per dono o per deposito legale e proprio su questo aspetto è necessario soffermarsi, dal momento che esso rappresenta un ulteriore elemento di criticità. La normativa vigente, definita dalla legge 15 aprile 2004, n. 106 e dal relativo regolamento di attuazione (D.P.R. 3 maggio 2006, n. 252), è stata elaborata in un contesto in cui il fenomeno dell'autopubblicazione non aveva ancora assunto le dimensioni attuali. Di conseguenza, il testo normativo non contiene riferimenti espliciti al self publishing, né fornisce indicazioni puntuali sulle modalità di adempimento degli obblighi di deposito per queste opere.

Nel quadro normativo attuale, le opere autopubblicate sono assimilate ai prodotti dell'editoria tradizionale e pertanto sono soggette agli obblighi previsti per le pubblicazioni a stampa: la eventuale assenza dell'ISBN non li esonera dai vincoli del deposito legale giacché questi sono legati alla diffusione e all'uso pubblico dell'opera.¹³ In particolare, se nel caso dei prodotti dell'editoria tradizionale la legge individua nell'editore il responsabile della pubblicazione e di conseguenza il soggetto obbligato al deposito, nel self publishing è l'autore stesso a essere considerato il soggetto responsabile della pubblicazione.¹⁴ Inoltre, le opere in regime di self publishing sono assimilate

¹² <https://www.isbn.it/>.

¹³ Legge 15 aprile 2004, n. 106, art. 1, c. 1: "Al fine di conservare la memoria della cultura e della vita sociale italiana sono oggetto di deposito obbligatorio, di seguito denominato "deposito legale", i documenti destinati all'uso pubblico e fruibili mediante la lettura, l'ascolto e la visione, qualunque sia il loro processo tecnico di produzione, di edizione o di diffusione, ivi compresi i documenti finalizzati alla fruizione da parte di portatori di handicap. [...] (art. 1, c. 3) I documenti destinati al deposito legale sono quelli prodotti totalmente o parzialmente in Italia, offerti in vendita o altrimenti distribuiti e comunque non diffusi in ambito esclusivamente privato; per quanto attiene ai documenti sonori e audiovisivi, sono destinati al deposito legale anche quelli distribuiti su licenza per il mercato italiano".

¹⁴ Il Servizio II della Direzione Generale Biblioteche e istituti culturali del Ministero della Cultura ha pubblicato sul suo

alle opere cartacee per le quali sono previsti dei parziali esoneri rispetto agli obblighi di deposito legale, che prevedono regolarmente il deposito di quattro copie per ogni titolo, due destinate all'archivio nazionale della produzione editoriale presso le Biblioteche nazionali centrali di Firenze e Roma e due destinate all'archivio regionale, costituito da istituti individuati su tutto il territorio nazionale, sia a livello regionale che provinciale. Nel caso delle opere autopubblicate, quindi, l'autore è tenuto al deposito di due copie soltanto per ogni titolo, una per l'archivio nazionale della produzione editoriale presso la Biblioteca nazionale centrale di Firenze, e una per l'istituto depositario individuato nella provincia in cui risiede l'autore,¹⁵ ma l'assenza del dato relativo al luogo di pubblicazione rende più complesso il processo di verifica degli adempimenti editoriali a cui sono tenuti gli istituti depositari (D.P.R. 03 maggio 2006, n. 252, art. 41). Nello stesso tempo l'aver individuato nella provincia di residenza dell'autore un criterio geografico di attribuzione della competenza territoriale a livello regionale ha consentito di superare le difficoltà interpretative sorte in un primo momento rispetto a tutti quei titoli autopubblicati con Amazon che spesso riportavano tra le note tipografiche la scritta "Printed in Poland" (Puglisi 2020, 609). Tuttavia, a differenza delle pubblicazioni tradizionali, non si applicano ad essi le soglie quantitative previste per la tiratura iniziale (se inferiore alle 200 copie), dal momento che le opere in regime di self publishing sono spesso realizzate tramite stampa *on demand* e in quantitativi estremamente ridotti.¹⁶ Nella pratica, nonostante tali indicazioni e nonostante molte piattaforme offrano tra i loro servizi a pagamento anche quello di deposito legale presso gli istituti individuati dalla legge, molti autori non sono pienamente consapevoli degli obblighi derivanti dalla normativa vigente e tendono a inviare spontaneamente le proprie opere a un numero più ampio di istituti, inclusi quelli non tenuti alla conservazione.¹⁷

Non è possibile ancora stimare quale sarà l'impatto che avrà sul deposito legale l'acquisizione e la conservazione a lungo termine dei libri autopubblicati in formato digitale: la gestione del deposito di questi documenti è rimasta uno dei punti più deboli della normativa, poiché non è ancora stato emanato lo specifico regolamento con cui dovranno essere definite le modalità di archiviazione e conservazione dei documenti diffusi tramite rete informatica e liberamente accessibili, in attuazione dell'articolo 37 del D.P.R. 03 maggio 2006, n. 252.¹⁸ Quando il deposito legale dei documenti

sito delle FAQ che forniscono alcune risposte agli interrogativi più spesso posti dagli autori di self publishing agli uffici che gestiscono il deposito legale: <https://biblioteche.cultura.gov.it/it/documenti/DepositoLegale/F.A.Q-DepositoLegale.pdf>.

¹⁵ L'elenco degli istituti depositari a livello regionale è individuato dai D. M. 28 dicembre 2007 e 10 dicembre 2009 (<https://biblioteche.cultura.gov.it/it/Attivita/deposito-legale/index.html>).

¹⁶ D.P.R. 3 maggio 2006, n. 252, art. 9: "Il soggetto obbligato consegna una sola copia per l'archivio nazionale e una sola copia per l'archivio regionale della produzione editoriale per le opere che abbiano una tiratura limitata non superiore ai 200 esemplari o un valore commerciale non inferiore a 15.000,00 euro per ciascun esemplare, fatta esclusione per la musica a stampa".

¹⁷ La Biblioteca nazionale centrale di Roma ha scelto di acquisire tali pubblicazioni come donazioni, senza procedere all'emissione della ricevuta di avvenuto deposito prevista dalla legge, distinguendo così chiaramente tra adempimento normativo e acquisizione volontaria.

¹⁸ D.P.R. 03 maggio 2006, n. 252, art 37, c. 1: "Le modalità di deposito dei documenti diffusi tramite rete informatica sono definite con successivo regolamento adottato ai sensi dell'articolo 5, comma 1, della legge 15 aprile 2004, n. 106, su proposta del Ministro per i beni e le attività culturali, di concerto con il Ministro delegato per l'innovazione e le tecnologie, sentite le associazioni di categoria interessate, nonché la Commissione per il deposito legale, di cui all'articolo 42 e il Comitato consultivo permanente per il diritto d'autore".

diffusi tramite rete informatica andrà a regime, attraverso procedure di raccolta automatica direttamente da siti web e piattaforme online, esso consentirà alle biblioteche depositarie di realizzare un controllo bibliografico molto più efficiente ed estensivo e di superare i problemi legati alla conservazione selettiva e a campione, spesso ipotizzata per questa tipologia di pubblicazioni, ma che, al di là del giudizio sulla loro qualità, costituiscono una testimonianza di una nuova epoca nel campo dell'editoria.

Conclusioni: il self publishing come fenomeno editoriale e sociale

Alla luce delle considerazioni emerse, il self publishing si configura non soltanto come un modello produttivo alternativo all'editoria tradizionale, ma anche come un fenomeno culturale complesso, capace di incidere in profondità sulle dinamiche della produzione, circolazione e conservazione del sapere scritto. La sua rilevanza non può essere compresa esclusivamente in termini quantitativi o tecnologici, ma va interpretata anche come espressione di trasformazioni sociali più ampie, che investono il rapporto tra autore, pubblico e istituzioni della conoscenza.

Il self publishing rappresenta una forma di “democratizzazione imperfetta” dell'accesso alla pubblicazione, nella quale la riduzione delle barriere tecniche ed economiche non coincide necessariamente con una reale parità di visibilità o di riconoscimento culturale (Schroff 2024). Se da un lato l'autopubblicazione consente a soggetti tradizionalmente esclusi dai circuiti editoriali di rendere pubblica la propria produzione, dall'altro essa tende a riprodurre nuove forme di disuguaglianza, legate alla capacità di orientarsi all'interno delle piattaforme, di investire risorse economiche e di acquisire competenze editoriali e promozionali. In questo senso, il self publishing non rappresenta semplicemente un'alternativa all'editoria tradizionale, ma un terreno ibrido in cui si ridefiniscono i confini tra professionalità e diletterismo, tra produzione culturale e consumo, tra visibilità e marginalità (Carbone 2015).

Dal punto di vista biblioteconomico, tali trasformazioni pongono interrogativi di grande rilievo. Le difficoltà di identificazione, descrizione e conservazione dei libri autopubblicati non sono soltanto problemi tecnici, ma riflettono la tensione tra modelli consolidati di controllo bibliografico e nuove forme di produzione editoriale che sfuggono ai criteri tradizionali di selezione e legittimazione. In questo contesto, il self publishing incide direttamente su una delle funzioni fondamentali delle biblioteche, ossia la costruzione di strumenti affidabili di accesso alla conoscenza registrata. Quando le pubblicazioni sfuggono ai circuiti editoriali le biblioteche incontrano maggiori difficoltà nel garantire quella rappresentazione ordinata e sistematica della produzione editoriale che costituisce uno dei presupposti della loro funzione conoscitiva.

Parallelamente, il self publishing sollecita una ridefinizione delle pratiche di selezione documentaria. Nel modello editoriale tradizionale, la presenza di un editore e l'inserimento nei circuiti distributivi costituiscono elementi che, pur non esauendo i criteri di valutazione bibliotecaria, operano come forme preliminari di legittimazione culturale. In un contesto in cui la produzione editoriale si sviluppa in larga parte al di fuori di questi circuiti, la capacità delle biblioteche di documentare in modo completo e sistematico la cultura contemporanea risulta inevitabilmente messa alla prova. Il rischio non è più, come in passato, quello della perdita di documenti rari o difficilmente accessibili, ma quello di una documentazione discontinua e frammentaria, determinata dalla difficoltà di

intercettare e trattare una produzione editoriale diffusa, decentralizzata e spesso priva di adeguati strumenti di identificazione.

In questo senso, il self publishing non rappresenta soltanto una nuova modalità di produzione libraria, ma costituisce un banco di prova per il ruolo stesso delle biblioteche nel contesto informativo contemporaneo. Esso rende evidente come le funzioni di controllo bibliografico, selezione e conservazione richiedano l'elaborazione di nuove strategie e modelli operativi, capaci di garantire la rappresentazione e la trasmissione della memoria culturale anche in un ecosistema editoriale caratterizzato da elevata frammentazione e da una progressiva disintermediazione.

In secondo luogo, risultano ridefinite le modalità attraverso cui le biblioteche esercitano la propria funzione di controllo bibliografico. La presenza di pubblicazioni prive di identificatori standard, di responsabilità editoriali chiaramente definite o di metadati affidabili rende più difficile garantire quella coerenza descrittiva e quella affidabilità informativa che costituiscono il presupposto per la condivisione delle registrazioni bibliografiche e per il funzionamento dei cataloghi collettivi. Sebbene gli standard catalografici consentano, in linea di principio, la descrizione di qualsiasi risorsa sulla base delle informazioni desumibili dalla pubblicazione stessa, la frequente assenza, nelle opere autopubblicate, di elementi presentati in forma chiara e strutturata rende più complessa la loro applicazione.

La questione assume una rilevanza ancora maggiore se considerata in relazione alle politiche di deposito legale e alla funzione di conservazione della memoria culturale. L'assenza di procedure pienamente operative per il deposito delle pubblicazioni digitali, unita alla frammentazione dei canali di produzione e distribuzione, rischia infatti di determinare lacune significative nella documentazione del presente.

In conclusione, l'interesse del self publishing non risiede esclusivamente nella sua dimensione quantitativa o economica, ma soprattutto nel suo valore sociologico e documentario. Esso costituisce una fonte privilegiata per osservare le trasformazioni delle pratiche autoriali, dei modelli di diffusione del sapere e delle forme di partecipazione culturale nella contemporaneità. Per le biblioteche, affrontare in modo sistematico le sfide poste dall'autopubblicazione significa non solo adeguare strumenti e procedure, ma anche riaffermare il proprio ruolo di garanti della memoria collettiva in un contesto editoriale sempre più frammentato, fluido e decentralizzato, in cui la presenza dei libri autopubblicati è ormai una componente strutturale della produzione libraria italiana che non può essere ignorata o peggio negata (un libro su sei pubblicato in Italia è un autopubblicato), essendo essi a tutti gli effetti prodotti del costume e della vita sociale di questi anni.

Riferimenti bibliografici*

AIB (Associazione italiana biblioteche). Commissione nazionale biblioteche e servizi nazionali, a c. di. 2019. "Il deposito legale regionale in Italia: stato dell'arte e risultati di una recente indagine." *AIB studi* 59: 423-52. <https://doi.org/10.2426/aibstudi-12019>.

AIE (Associazione Italiana Editori). 2024. *Rapporto sullo stato dell'editoria in Italia 2024 su consolidato 2023*, a cura dell'Ufficio Studi AIE. Milano: Associazione Italiana Editori.

Carbone, Andrea Libero. 2015 "Il self-publishing non esiste." *Doppiozero*. <https://www.doppiozero.com/il-self-publishing-non-esiste>.

Dinotola, Sara. 2021. "Offerta editoriale e collezioni bibliotecarie: uno studio comparato." *AIB studi* 61: 387-424. <https://doi.org/10.2426/aibstudi-13216>.

Garzia, Mario. 2023. "Il self publishing in Italia: dall'autopubblicazione all'autoeditoria. La sfida della disintermediazione." In *Verso l'editoria digitale: storia, innovazioni e ibridazioni del sistema editoriale in Italia*, a cura di Emiliano Ilardi e Giovanni Ragone, 49-75. Napoli: Liguori.

O'Dair, Ellen. 2018. "Acquiring self-published material: challenges and opportunities for libraries." *Moving through the grey: publishing in action* 1: 145-52. <https://ojs.scholarsportal.info/ottawa-school/index.php/PA/article/view/144>.

Holley, Robert P. 2014. *Self-Publishing: a new challenge for Universal Bibliographic Control*. <https://repository.ifla.org/items/86748e9e-69eb-4265-b20a-e5348f854a5d>.

Manzoni, Laura. 2023. "CIP e metadattazione condivisa nell'ambiente digitale come nuove forme di cooperazione tra editori e biblioteche." *Sistema editoria* 2: 57-76. <https://doi.org/10.14672/se.v1i2.2241>.

Moretti, Simone. 2013. *Il self publishing in Italia*. <https://simonemoretti75.wordpress.com/2013/11/28/il-self-publishing-in-italia/>.

Pacer, Matt. 2013. "Small press and self-publishing books: a collection development dilemma." *Against the grain* 25: 33-4. <https://doi.org/10.7771/2380-176X.6521>.

Peresson, Giovanni. 2016. *Il mercato del self-publishing in Italia*. <https://plpl.it/news/2016/il-mercato-del-self-publishing-in-italia/>.

Puglisi, Paola. 2020. "Deposito legale quattordici anni dopo: come, quando, "quanto", e perché." *AIB studi* 60: 591-614. <https://doi.org/10.2426/aibstudi-12477>.

Ragone, Giovanni. 2005. *L'editoria in Italia: storia e scenari per il XXI secolo*. Napoli: Liguori.

Ragone, Giovanni. 2023. "L'editoria di quarta generazione." In *Verso l'editoria digitale: storia, innovazioni e ibridazioni del sistema editoriale in Italia*, a cura di Emiliano Ilardi e Giovanni Ragone, 1-12. Napoli: Liguori.

Schroff, Simone. 2024. "Missing the point: the disappointed hope of self-publishing authors." *International journal of cultural policy* 30: 960-77. <https://doi.org/10.1080/10286632.2023.2288584>.

* Ultima data di consultazione dei siti web: 22 gennaio 2026.

Tuncer, Nurhak, e David Reed, 2019. “The cataloguing of self-published items.” *Cataloging & Classification Quarterly* 57: 206–26.

Zita, Tiziana. 2016. *La straordinaria crescita del self publishing in Italia*. <https://www.cronacheletterarie.com/2016/12/15/il-mercato-del-self-publishing-in-italia/>.

Institutional practices in gender recording: an analysis of National Libraries*

Ana Carolina Novaes de Mendonça^(a)
Ana Carolina Simionato Arakaki^(b)

a) Federal University of São Carlos and Federal District Research Support Foundation, <https://orcid.org/0009-0004-0285-9932>

b) Faculty of Information Science, University of Brasília; Federal University of São Carlos and Federal District Research Support Foundation, <https://orcid.org/0000-0002-0140-9110>

Contact: Ana Carolina Novaes de Mendonça, mnovaesana@gmail.com;
Ana Carolina Simionato Arakaki, ana.arakaki@unb.br

Received: 08 July 2025; **Accepted:** 16 March 2026; **First Published:** 15 May 2026

ABSTRACT

The representation of gender metadata in personal name authority records maintained by national libraries in the Americas and Europe is the central focus of this study. Anchored in contemporary discussions on descriptive practices and standards, the research investigates the application of frameworks such as RDA, alongside metadata encoding formats like MARC21, UNIMARC, and INTERMARC, in the inclusion and treatment of gender-related information. Adopting a descriptive and exploratory methodology, the study examines public authority catalogues and analyses institutional responses to identify how gender is recorded, made visible, or omitted. The findings reveal substantial variation across institutions: while some adopt binary classification models, others employ more inclusive and expanded vocabularies, and certain libraries opt to suppress gender data from public access. The analysis further highlights the potential of semantic web technologies and linked data as ethical and sustainable alternatives, enabling the connection of authority records to external sources such as VIAF, Wikidata, and ORCID. The study concludes that, when managed with responsibility and transparency, gender metadata not only supports academic research and improves data quality but also enhances the visibility of diverse contributions to cultural and scientific production.

KEYWORDS

Name Authority records; Gender; Linked Data; Semantic Web.

* Ana Carolina Novaes de Mendonça and Ana Carolina Simionato Arakaki are the authors of sections 1, 2, 3, 4, and the conclusion. Data collection was carried out by Ana Carolina Novaes de Mendonça.

We gratefully acknowledge the support of the National Council for Scientific and Technological Development (CNPq), the Brazilian Institute of Information in Science and Technology (IBICT), and the Research Support Foundation of the Federal District (FAPDF).

© 2026, The Author(s). This is an open access article, free of all copyright, that anyone can freely read, download, copy, distribute, print, search, or link to the full texts or use them for any other lawful purpose. This article is made available under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. JLIS.it is a journal of the SAGAS Department, University of Florence, Italy, published by EUM, Edizioni Università di Macerata, Italy, and FUP, Firenze University Press, Italy.

1. Introduction

In recent years, the growing awareness of gender equality has extended to the field of Information Science, aligning with global initiatives such as the United Nations Sustainable Development Goals (SDGs), particularly Goal 5 (United Nations 2015). Within the field of cataloguing, we consider the possibility that adequate representation of gender in authority records can contribute to inclusion, visibility, and improved information retrieval.

Authority control systems, in their more recent conceptualization, are intended to support the identification and disambiguation of entities. These systems operate under institutional policies while relying on shared principles of standardization and interoperability. Descriptive guidelines such as the Anglo-American Cataloguing Rules (AACR2r) and Resource Description and Access (RDA) provide structured criteria for bibliographic description. RDA offers greater flexibility and alignment with conceptual models, including the optional inclusion of attributes such as gender. In metadata schemas, MARC21 uses field 375 to record gender information, while INTERMARC employs field 120 for the same purpose. However, the specific values used to represent gender vary across institutions, depending on local cataloguing policies, controlled vocabularies, and the extent of alignment with international standards.

The inclusion of gender in personal name authority records (NARs) has raised ethical concerns related to privacy, consent, and the risk of misrepresentation, especially for transgender and non-binary individuals. This debate has led institutions to move away from recording gender as a fixed attribute, reinforcing the need for alternatives that respect identities while ensuring data integrity and interoperability. In this scenario, the Semantic Web and Linked Data offer flexible, decentralized, and ethically aligned solutions for gender representation. Semantic enrichment has become a key strategy for making metadata more granular, multilingual, and semantically meaningful, particularly in Libraries, Archives, and Museums (LAMs). As Zeng (2019) notes, this process improves data quality, discoverability, and contextualization. This approach is increasingly relevant to authority records for representing identity attributes more inclusively.

Although this study does not focus on the conceptual debate surrounding gender, it adopts the understanding that gender is understood as a form of self-identification. According to Butler (1990), gender does not constitute a fixed identity, but rather a set of repeated and socially regulated practices. This conception questions the validity of authority systems that operate with binary and static categories. As Scott (1995) posits, gender is a social construct that transcends biological sex, encompassing identities, behaviours, social expectations, and roles.

Grounded in this understanding, the present study shifts from theoretical discussions to the practical and technical dimensions of gender representation in authority data. The epistemology underlying bibliographic authority systems often operates with normative assumptions about identity. These systems tend to establish fixed categories based on criteria external to individuals' self-identification. These assumptions are challenged by gender studies, which propose a relational and situated ontology and emphasize the plurality and fluidity of identities. In this context, the study addresses the following research question: how do national libraries represent gender metadata in personal name authority records, and what institutional policies guide these practices? The objective is to examine how national libraries include gender metadata in authority records. The study focuses on institutional decisions guiding these practices and explores the potential use of semantic enrichment strategies to improve interoperability and inclusivity.

2. Theoretical framework

With the exponential growth of data, effective representation becomes essential. Marcondes (2001) argues that information is only useful if it can be found, underscoring the importance of discoverability. Bibliographic and authority records result from representational choices that shape visibility in catalogues. The cataloguer's judgment directly influences how information is organized and accessed, potentially highlighting or obscuring certain individuals, groups, or topics (Martin 2021).

Olson (2002) highlights that knowledge organization systems often reflect dominant power structures, making identities that fall outside traditional norms invisible. This perspective is key to understanding the epistemological challenges of including gender in standardized authority records. Ethical and technical guidelines, such as the Cataloguing Code of Ethics (2021) and the IFLA's International Cataloguing Principles (2016), aim to promote fairness, accuracy, and consistency in cataloguing practices. Authority records play a central role in this context by documenting decisions and connecting name variants, functioning both as technical mechanisms and intellectual frameworks (Taylor and Joudrey 2009; Tillett 2004).

As Possemato (2025) observes, conceptual models have shifted cataloguing practices toward a more contextualized description of entities. Oliver (2021) further emphasizes that the RDA was designed for use beyond traditional library settings, with the goal of supporting Linked Data principles and encouraging the adoption of Semantic Web technologies to ensure the continued relevance of bibliographic infrastructure.

The inclusion of gender as an attribute in personal name authority records (NARs) is conceptually grounded in the FRAD, which informs the structure and logic of the RDA standard. In this framework, RDA instruction 9.7 defines gender as an optional descriptive element and recommends the use of controlled vocabulary terms such as "male," "female," or "unknown," allowing cataloguers to include this attribute when relevant and in accordance with institutional policies. Building on this shared foundation, the creation, exchange, and long-term management of authority records depend on the use of standardized metadata schemas, which are essential for maintaining consistency and interoperability across institutions and platforms. MARC 21 has long been the primary format for encoding bibliographic and authority data, while newer models like BIBFRAME align with Linked Data and Semantic Web principles. Among MARC 21's five formats, the Authority format standardizes names, titles, and subjects. Following the development of RDA, updates to MARC 21 such as the addition of 3XX fields have enhanced the descriptive detail of entities (Seikel and Steele 2011).

These fields were designed to enhance the description of entities, enabling more granular identification of persons, corporate bodies, and works. Within this structure, field 375 was incorporated to record gender as a personal attribute. According to its scope note, this field specifies "information about a person's gender, including applicable dates" (Library of Congress 2022). Although optional, it provides a structured way to represent gender identity in authority records. Field 375 includes subfield \$a, which records the gender with which the person identifies.¹

¹ Other subfields complement the 375 structure: \$s and \$t indicate the start and end dates of the identification, when applicable; \$u links to an external URI for the gender term; \$v indicates the source of the information; \$0 contains an authority

Since 2001, the UNIMARC Authority format has included field 120 to record gender, using pre-defined values: “a” (female), “b” (male), “c” (transgender), “u” (unknown), and “x” (not applicable).² In this context, the inclusion of gender in authority records has become a prominent topic of discussion within cataloguing communities.³

The Library of Congress and the Name Authority Cooperative Program (NACO) instructed cataloguers to include gender in authority records based on a binary model, often relying on assumptions (Martin 2021). This practice has been widely criticized by Billey, Drabinski, and Roberto (2014), who, grounded in queer theory, argue that gender is a socially constructed category and condemn the reinforcement of fixed, binary identities. They also highlight the risks of inferring gender from names or biographical data, stressing that retrieval functionalities should not outweigh the need for accurate and respectful representation. Amid these debates, the PCC Ad Hoc Task Group on Gender in Name Authority Records recommended in 2016 that gender be recorded only upon explicit self-identification. In 2022, the group advised discontinuing the use of field 375 and removing it from existing records, allowing gender information only in unstructured fields like 670 or 678 (PCC Ad Hoc Task Group on Recording Gender in Personal Name Authority Records 2022).

Thompson (2016) examines Name Authority Records (NARs) related to trans authors, highlighting how cataloguing practices based on the FRAD and RDA models particularly the use of field 375 for recording gender can result in involuntary outing and violations of privacy. As an alternative, the author proposes that authors be granted greater agency to self-identify through systems such as ORCID or linked data platforms and advocates for a more ethical cataloguing approach, grounded in privacy, consent, and respect for self-identification, to avoid discriminatory practices and promote fairer representation.

Adolpho (2019) argues that the rigidity of cataloguing models limits the adequate representation of gender identity. Imposing fixed notions of gender fails to reflect its diversity and risks reinforcing exclusion and discrimination. To respect self-determination, the author advocates abandoning binary and inflexible models in favour of consent-based and adaptable practices. Burns et al. (2022) align with PCC guidance in treating gender as a high-risk attribute in authority data: it can be inaccurate, misrepresent trans people, and enable non-consensual disclosure. The report therefore recommends not recording gender in authority records unless a clear purpose exists and informed consent is obtained, prioritising self-identification, suppression of former names when requested, and the use of inclusive vocabularies such as Homosaurus.⁴

control number or identifier; \$1 provides a persistent URI for a real-world entity; \$2 identifies the source vocabulary, such as the Gender, Sex, and Sexual Orientation (GSSO) ontology; \$6 links language versions of the record; \$7 captures provenance and reliability; and \$8 manages relationships between fields within the same record (Library of Congress 2022).

² In 2023, the online edition of the UNIMARC Manual for Authorities introduced updates aligned with the IFLA-LRM, restructuring fields and subfields to support both traditional bibliographic practices and models based on entities and relationships (IFLA 2024).

³ Several studies and institutional initiatives have addressed the implications, challenges, and possibilities surrounding this practice, including works by Billey, Drabinski, and Roberto (2014); Thompson (2016); Adolpho (2019); Wainer (2019); the Program for Cooperative Cataloguing (PCC) on Gender in Name Authority Records (2016, 2022); Carbajal (2020); Martin (2021); Burns et al. (2022); Lindsey, Suiter, and Hanselman (2022); NISO (2022) and Possemato (2025).

⁴ <https://homosaurus.org/>

This debate remains ongoing and continues to evolve. As Possemato (2025) observes, the inclusion of gender in authority records takes place within a broader movement of transformation in cataloguing practices, driven by the adoption of Semantic Web technologies and the transition toward entity-based models. The traditional view, which regarded gender as a fixed and inherent attribute of the entity, is increasingly being replaced by approaches that acknowledge the sociocultural construction and fluidity of gender. Also, although often criticized for reinforcing binary norms, the inclusion of gender information in authority records can offer meaningful benefits, such as enhancing the visibility of marginalized identities, improving information retrieval, and supporting the needs of specific communities. However, these benefits are only realized when such metadata is recorded with informed consent, grounded in self-identification, and allows for future revision or suppression. As emphasized in the Metadata Best Practices for Trans and Gender Diverse Resources, “gender information can enhance discoverability, support community needs, and affirm the identities of individuals or creators who choose to have this metadata included” (Gender & Sexuality in Information Studies Special Interest Group 2022, 4). Billey, Drabinski, and Roberto (2014) argue that classification and subject access systems are not neutral. They embed dominant ideologies that can marginalize queer and nonbinary identities. Also warns that metadata structures often fail to reflect the complexities of transgender experiences, particularly when they rely on rigid authority control mechanisms that disregard the nuances of gender fluidity. Together, these perspectives underscore the ethical imperative of designing metadata practices that affirm identity while resisting the enforcement of normative taxonomies.

2.1 Semantic Approaches to Gender Representation in Personal Name Authority Records

The Semantic Web, proposed by Berners-Lee, Hendler, and Lassila (2001), addresses the Web’s limitation in enabling machines to interpret and process data meaning. As Pomerantz (2015) notes, it is not just about publishing data online but about structuring it in an interconnected way. In this context, Berners-Lee (2006) introduced Linked Data as a set of practices for publishing and linking data in machine-readable formats. Institutions like the Library of Congress, the National Library of France (BNF), the German National Library (DNB), and the National Library of Spain (BNE) apply these principles by publishing authority data in RDF and offering SPARQL endpoints. These implementations demonstrate how authority records can transition from rigid cataloguing frameworks to more dynamic, semantically enriched environments.

Semantic enrichment involves generating new metadata or enhancing existing ones by adding contextual and meaningful information, improving interoperability and connection to external datasets (Isaac et al. 2015). This process goes beyond adding data, as it establishes semantic relationships that situate objects within broader contexts, linking them to entities like people, places, or topics (Manguinhas et al. 2016).

Linked Data sources provide external identifiers and relationships that can be used in authority enrichment. According to Smith-Yoshimura (2018), the most used are id.loc.gov, VIAF, DBpedia, GeoNames, and Wikidata, with notable growth in Wikidata between 2015 and 2018. VIAF presents gender as textual values aggregated from institutions, while Wikidata encodes gender using

persistent URIs, enabling semantic interoperability. Semantic enrichment connects authority records to external sources like VIAF, Wikidata, DBpedia, and GSSO, incorporating diverse and community-driven definitions to improve accuracy and ethical representation. For gender representation, this approach may offer an alternative to embedding gender directly in authority records. By referencing public data from reliable sources with documented provenance and update histories, institutions can support more transparent maintenance of gender-related statements and other attributes over time.

Thompson (2016) reinforces this perspective by highlighting that recording gender within authority records particularly through MARC field 375 poses significant risks related to privacy breaches, non-consensual disclosure, and the reinforcement of cis-normative frameworks. She advocates for a shift toward linked data practices, where identity information is managed through persistent identifiers, allowing individuals greater control over how they are represented. Systems like ORCID exemplify this model, offering users the ability to manage their own identity metadata with granular privacy settings. Also argues that cataloguers should critically assess whether including gender is truly necessary for information retrieval, always prioritizing consent and the prevention of harm.

The adoption of semantic web technologies and semantic enrichment practices can be a technical advancement for authority records and an ethical transformation in managing identity attributes, particularly gender. This moves authority control from rigid, closed metadata models to open, flexible, and community-driven frameworks that support greater inclusivity, interoperability, and respect for self-identification. By leveraging linked data sources such as Wikidata, VIAF, and ORCID, may offer pathways for self-identification and ongoing updates. This approach aligns with broader changes in bibliographic control, fostering transparency, autonomy, and sustainability while reducing the risks of harm, misrepresentation, and the reinforcement of normative biases in cataloguing systems.

3. Methodology

From the perspective of its objective, this study is classified as descriptive and exploratory, as it seeks to identify, characterize, and analyze how gender metadata is implemented in national libraries' authority records. According to Gil (2008), descriptive research aims to identify and describe the characteristics of a given phenomenon or population, while exploratory research seeks to provide greater familiarity with a problem, making it clearer and enabling the formulation of hypotheses (Prodanov and Freitas 2013). This combined approach enables the investigation of gender representation practices from multiple angles. In terms of its nature, the study adopts an analytical-descriptive orientation, grounded in empirical observation and institutional documentation.

The study employs a qualitative approach and uses bibliographic and documentary research methods. Bibliographic research is based on the analysis of previously published materials, such as books and academic articles. Documentary research, in turn, involves diverse and dispersed sources, including institutional reports, technical standards, and official documents that may not yet have undergone in-depth analytical treatment. As Gil (2008) highlights, both strategies con-

tribute to understanding the current state of knowledge on the topic and provide a solid theoretical foundation.

The inclusion criteria considered studies available in full text, published in Portuguese, English, or Spanish, between 2000 and 2024, to ensure the incorporation of contemporary and relevant research. To guarantee the breadth and quality of the selected literature, databases widely recognized in the field of Library and Information Science were consulted, including the Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (BRAPCI),⁵ the CAPES Portal of Journals,⁶ Library and Information Science Abstracts (LISA),⁷ Web of Science,⁸ Scopus,⁹ and the Scientific Electronic Library Online (SciELO).¹⁰

Based on the descriptive and exploratory nature of the study, national libraries from the Americas and Europe were identified using the continental classification available on IFLA National Libraries database and Wikipedia¹¹ as an initial reference. Data collection involved the analysis of personal name authority records available in public library catalogues, complemented by direct contact via email with the respective institutions. The purpose of this contact was to verify whether gender metadata was included in their authority records, to understand whether they used specific fields for recording gender, and how this information was being registered. Formal emails were sent to the selected institutions to request official information about their policies and practices regarding the inclusion of gender metadata in authority records.

In the empirical analysis, one authority record was examined for each institution. The selected records correspond to the authors Agatha Christie, J. L. Butler, Jane Austen, Lola Anglada, and Virginia Woolf, chosen due to their wide presence across authority databases and their representativeness in diverse historical, cultural, and linguistic contexts. In total, thirteen records from thirteen national libraries were analyzed, corresponding to the institutions that confirmed the inclusion of gender metadata in their authority records. The aim was to obtain institutional positions, not personal opinions. The responses were treated as official statements, providing complementary insights to the analysis of public authority catalogues.

From the thirteen libraries identified, four authority catalogues were randomly selected to illustrate how gender metadata is represented in practice. The selection was made randomly among these institutions.

4. Results and discussion

After completing the data collection and contacting the national libraries, the institutions that confirmed the use of gender metadata in their authority records were organized in Table 1. The

⁵ <https://brapci.inf.br/home>

⁶ <https://www.periodicos.capes.gov.br/>

⁷ <https://about.proquest.com/en/products-services/lisa-set-c>

⁸ <https://www.webofscience.com/wos/woscc/basic-search>

⁹ <https://www.scopus.com/sources>

¹⁰ <https://www.scielo.br/>

¹¹ Classification of National Libraries on Wikipedia: https://pt.wikipedia.org/wiki/Biblioteca_nacional.

table was divided into two groups: national libraries from the Americas and from Europe.¹² The practices adopted, considering three main elements: the field and subfield used to record gender; the values adopted to represent this information; and the institutional decisions that guide or justify the use of this metadata. The systematization of this information enables an understanding of how metadata standards and descriptive schemas are interpreted and applied by each library, as well as the degree of alignment or adaptation to international guidelines, local policies, and ongoing ethical discussions surrounding the representation of gender identity in authority records.

Table 1. Gender Metadata Implementation in Authority Records of National Libraries

AMERICAS				
Library	Country	Field and sub-field used	Values	Practice on gender recording
National Library Mariano Moreno	Argentina	MARC 21 field 375 Subfield \$a	Woman Man Trans woman Trans man Intersex Travesti Non-binary Other Unspecified	Records gender and has expanded the set of values.
National Library of Colombia	Colombia	MARC 21 field 375 Subfield \$a	Female Male Unknown	Implemented, in 2021, the use of MARC 21 field 375 to record the gender of authors in authority records, following the guidelines established in RDA.
National Library of Mexico	Mexico	MARC 21 field 375 Subfield \$a	Female Male	Gender is recorded as either female or male, and the “unknown” option is not used. In cases where the author identifies with another gender, a non-controlled or transcribed term from the consulted source is recorded.
National Library of Peru	Peru	MARC 21 field 375 Subfield \$a	Female Male	Follows the guidelines of the Manual for the control of authority data for people, based on RDA. Although there is a policy supporting the identification, retrieval, and gender grouping of authors, internal policy determines that this information is not publicly displayed in the authority catalogue.

¹² The American group comprises the National Library of Argentina (Biblioteca Nacional de la República Argentina), the National Library of Colombia (Biblioteca Nacional de Colombia), the National Library of Mexico (Biblioteca Nacional de México), and the National Library of Peru (Biblioteca Nacional de Perú). The European group includes the National Library of France (Bibliothèque Nationale de France – BnF), the Royal Library of Belgium, the National Library of Iceland (Landsbókasafn Íslands – Háskólabókasafn), the National Library of Norway (Nasjonalbiblioteket), the National Library of the Czech Republic (Národní knihovna České republiky), the National Library of Spain (Biblioteca Nacional de España – BNE) and the German National Library (Deutsche Nationalbibliothek – DNB).

EUROPE				
Library	Country	Field and sub-field used	Values	Practice on gender recording
National Library of Germany	Germany	MARC 21 field 375 Subfield \$a	F (female) M (male)	Uses only the ISO 5218 codes: “m” (male) and “f” (female), with no specific option for other gender identities, as codes “0” (unknown) and “9” (not applicable) are not considered appropriate.
Royal Library of Belgium	Belgium	MARC 21 field 375 Subfield \$a	Female Male Unknown	Follows the guidelines established in the RDA.
National Library of Catalonia	Spain – Catalonia	MARC 21 field 375 Subfield \$a	Female Male	Follows the RDA guidelines. In CANTIC, this field is optional and depends on the policy of the contributing institution. There is no standardized list of gender identities. When necessary, a term from a controlled vocabulary may be used.
National Library of Spain	Spain	MARC 21 field 375 Subfield \$a	Female Male	Uses field 375 with the values “Male” and “Female” (other values are not applied).
National Library of Slovenia	Slovenia	COMARC/A field 120 subfield \$a	a (female) b (male) c (transgender) n (unspecified or not determined)	In the COBISS system, the COMARC/A format is used for authority records, based on UNIMARC/A. Gender is recorded in field 120\$a.
National Library of France	France	INTERMARC field 008 (position 17) UNIMARC field 120 subfield \$a	F (female) M (male) # (not informed) a (female) b (male)	Data is produced in INTERMARC and also disclosed in UNIMARC. In INTERMARC, gender is recorded in position 17 of field 008, using the values: <i>f</i> (female), <i>m</i> (male), and # (not informed). In UNIMARC, the information appears in field 120 \$a.
National Library of Iceland	Iceland	MARC 21 field 375 Subfield \$a	Female Male Other	The values used are “male,” “female,” and “other.” If there is uncertainty about which value to use, the field is not included in the authority record.
National Library of Norway	Norway	MARC 21 field 375 Subfield \$a	F (female) M (male)	Accepted values for subfield \$a are “f” (female) and “m” (male). The field is repeatable and, in cases of gender transition, subfields \$s (start date) and \$t (end date) can be used to indicate the period of identification with a specific gender.
National Library of the Czech Republic	Czech Republic	MARC 21 field 375 Subfield \$a	Female Male	The values are “Male” and “Female.” Other values are not contemplated.

The National Libraries of Czech Republic,¹³ Mexico, Norway, Peru, and Spain follow similar practices by using MARC 21 field 375, subfield \$a, restricted to the binary values “Female” and “Male”. Mexico does not use the “Unknown” value and allows cataloguers to add a non-controlled or transcribed term when the person identifies outside the binary. In Norway, the field is repeatable and allows the use of subfields \$s (start date) and \$t (end date) to indicate periods of gender identification, especially in cases of gender transition. In Peru, although gender is recorded, the information is not displayed publicly due to internal policies focused on privacy and data ethics. The National Libraries of Colombia and Belgium apply a slightly broader model. Both are aligned with RDA guidelines, using the values “Female”, “Male”, and “Unknown”.

The National Library of Iceland adopts a more open approach, using MARC 21, field 375, subfield \$a, with the values “Female”, “Male”, and “Other”. When uncertainty exists regarding which value to assign, the field is left blank.

The National Library of Catalonia¹⁴ applies the same field and subfield. In the CANTIC cooperative catalogue, this field is optional and depends on the policy of each contributing institution. There is no standardized list of gender identities, however, when needed, terms from controlled vocabularies may be applied.

The National Library Mariano Moreno (Argentina) demonstrates one of the most comprehensive and inclusive approaches. It uses MARC 21 field 375, subfield \$a, and adopts an expanded set of values beyond the binary. This practice reflects an institutional decision¹⁵ to promote diversity and visibility for gender-diverse identities within authority data.

National Library of Germany uses MARC 21, field 375, subfield \$a, with ISO 5218 codes “m” (male) and “f” (female). There is no option for identities beyond the binary, as codes “0” (unknown) and “9” (not applicable) are considered inappropriate and are not used.

In contrast, the National Library of Slovenia, which adopts the COMARC/A format, records gender in field 120, subfield \$a, with the values “a” (female), “b” (male), “c” (transgender), and “n” (unspecified or not determined).

Finally, the National Library of France uses both INTERMARC and UNIMARC. In INTERMARC, gender is recorded in position 17 of field 008 with values “f” (female), “m” (male), and “#” (not informed). In UNIMARC,¹⁶ the same information appears in field 120, subfield \$a, using “a” (female) and “b” (male).

This analysis reveals that the inclusion of gender metadata in authority records is not simply a technical decision but one that reflects deeper institutional values, ethical considerations, and differing interpretations of metadata standards. While some libraries adopt a strictly binary framework, others demonstrate incremental steps toward more inclusive models.

To illustrate how gender metadata has been applied in authority records, the following section presents examples of records for the author Jane Austen in selected national library catalogues.

¹³ Institutional practice from the National Library of Czech Republic: <https://authority.nkp.cz/jmenne-autority/metodicka-materialy/metodika-jmena-cvicne-2#pole375>.

¹⁴ Institutional practice from the National Library of Catalonia: <https://www.bnc.cat/sprof/normalitzacio/marc21a/a1xx.htm#Auto375>.

¹⁵ Institutional practice from the National Library Mariano Moreno: <https://escritoriop.bn.gov.ar/mcat/mcata-3XX.html>.

¹⁶ Institutional practice from the National Library of France: <https://www.transition-bibliographique.fr/unimarc/formats-unimarc/>.

The choice of this author is based on her widespread presence across various authority databases, as well as her significance in diverse historical, cultural, and linguistic contexts.

4.1 Gender Metadata in Personal Name Authority Records

The National Library Mariano Moreno (BNMM) of Argentina uses the MARC 21 format in its authority records. Figure 1 shows the authority record for the author Jane Austen, where the presence of field 375, designated for recording gender, can be observed. This field reflects the institution's policy of adopting expanded and more inclusive vocabularies to represent gender identities, aligning with practices that promote greater diversity and visibility in bibliographic representation.¹⁷

FMT	AU
LDR	cz a22 n 4500
001	000021711
003	AR-BaBN
005	20221201132756.0
008	790418nj acannaabn lb aaa
0247	1a 102333412 2 viaf 1 http://viaf.org/viaf/102333412
040	1a DLC b eng c DLC d DLC d CIY-BR d OCoLC d DLC-R d SG-SIILA d OCoLC d DLC d CST d UPB d AR-BaBN
043	1a e-uk-en
046	f 17751216 g 18170718
1001	1a Austen, Jane, d 1775-1817
370	1a Steventon (Hampshire, Inglaterra) b Winchester (Inglaterra) e Bath (Inglaterra)
374	1a Autores
375	1a Mujer ←
377	1a eng l inglés
4001	1a Ostin, Dzhein, d 1775-1817
4000	1a Ao-ssu-ting, d 1775-1817
4000	1a Aosiding, d 1775-1817
4001	1a Ao-ssu-ting, Chien, d 1775-1817
4001	1a Aosiding, Jian, d 1775-1817
4001	1a Austenová , Jane, d 1775-1817
670	1a Wikipedia, via WWW, Feb. 8, 2012: b (n. dic. 16, 1775 en Steventon, Hampshire; m. 18 jul. 1817 en Winchester, Hampshire; novelista inglesa; 1800 se traslada a Bath)
670	1a Persuación, 1978.
6780	1a Nació en 1775 y falleció en 1817, en su hogar contaba con una biblioteca privada que la sumergió en la lectura, su talento como escritora aflora en su adolescencia, fué una destacada novelista británica. Algunas de sus obras son llevadas al cine. La mayoría de sus obras aparecen bajo seudónimo.

Figure 1. Authority record for the author Jane Austen in the BNMM

¹⁷ The BNMM adopted fields 370, 372, 373, 374, 375, and 377 to enrich personal name authority records. Initially, field 375 was limited to the values “Masculino” (Male) and “Femenino” (Female), based on ISO standards. During the pandemic, the institution launched a campaign to collect biographical data directly from authors, prioritizing self-reported information. Guided by the policies of Argentina’s Ministry of Women, the Gender Commission proposed expanding the terms used for gender. Despite the RDA working groups recommending the removal of field 375 due to privacy concerns, the BNMM maintains its use based on Argentina’s Personal Data Protection Law, which permits the processing of this data when authorized by the individual or obtained from public sources. The Authority Control Division argues that including gender metadata supports the analysis of bibliographic production by gender and fosters new research perspectives.

Among the various fields displayed, field 375 subfield \$a stands out, indicating the gender of the described entity with the value “Mujer” (Female). Field 100 presents the authorized name; 370 records the places associated with the author; and field 374 indicates her occupation. Additionally, field 377 specifies the language associated, while fields 400 and 500 list variants of the author’s name. Field 670 contains external references, including Wikipedia, and field 678 holds a biographical note highlighting Jane Austen’s relevance as a writer.

It is also possible to observe the application of semantic techniques that contribute to the interoperability of authority data. A key indicator of this practice is field 0247, which contains a VIAF identifier linked to the described entity. This field establishes a semantic link between the local authority record and an international authority base, allowing interoperable systems to recognize and reuse the information automatically.

4.2 German National Library

Figure 2 presents the authority record for Jane Austen in the catalogue of the German National Library (DNB), within the Gemeinsame Normdatei (GND). According to the information provided by the library, the use of field 375 is optional. The field “Geschlecht” (Gender) is populated with “weiblich” (female), demonstrating the library’s practice of recording this information.

Link zu diesem Datensatz	https://d-nb.info/gnd/118505173
Person	Austen, Jane
Geschlecht	weiblich 
Andere Namen	Aosiding (Chines.) Aosiding, Jian Ao-ssu-ting, Chien Ao-ssu-ting Austenová, Jane Austenové, Jane Austenová Jane Awsten, Jeyn (Uigur.) Osten, Dzhein Остен, Жейн Ostin, Dzhein (Russ. Namensform, AACR) Ostin, Džejn (Russ. Namensform, RAK-WB) Ostin, Zhejn Остин, Жейн Ostjen, Zhjejn 奧斯汀 (Schriftcode: Hans) 奧斯婷 (Chinesische Namensform (Transkription um 1950)) (Schriftcode: Hant) 珍·奧斯婷 (Chinesische Namensform (Transkription um 1950)) (Schriftcode: Hant) 简·奧斯汀 (Chinesische Namensform (Transkription wikipedia, Zugriff: 18.05.2021)) (Schriftcode: Hans) 簡·奧斯汀 (Chinesische Namensform (Transkription wikipedia, Zugriff: 18.05.2021)) (Schriftcode: Hant)
Quelle	M B 1986 Kröner Weltlit. LCAuth Wikipedia (Stand: 09.03.2023): https://de.wikipedia.org/w/index.php?title=Jane_Austen&oldid=231601159
Zeit	Lebensdaten: 1775-1817
Land	Großbritannien (XA-GB)
Sprache(n)	Englisch (eng)
Geografischer Bezug	Geburtsort: Steventon (Hampshire) Sterbeort: Winchester Wirkungsort: Bath
Beruf(e)	Schriftstellerin Librettistin

Figure 2. Authority record for the author Jane Austen in the DNB

The record also includes variant names derived from different writing systems and transliterations, including Chinese and Russian forms, along with biographical information such as dates, places of birth and death, associated language and occupation.

The library's catalogue incorporates Semantic Web technologies, allowing records to be exported in RDF Turtle format. This approach integrates authority data into the Linked Data environment, promoting interoperability and connectivity across various data sources. Such implementation reflects how authority catalogues are adapting to semantic environments, enhancing both the representation and retrieval of more contextualized information. Part of the same record, encoded in MARC 21 XML, is shown in Figure 3, highlighting how the data is structured according to this standard.

```
<datafield tag="100" ind1="1" ind2=" " >
  <subfield code="a">Austen, Jane</subfield>
  <subfield code="d">1775-1817</subfield>
</datafield>
<datafield tag="375" ind1=" " ind2=" " >
  <subfield code="a">2</subfield>
  <subfield code="2">iso5218</subfield>
</datafield>
```

Figure 3. Authority record in MARCXML for the author Jane Austen in the DNB

As shown in Figure 3, field 375 is encoded with the value “2” in subfield \$a and linked to the source vocabulary in subfield \$2 labeled “iso5218,” indicating compliance with ISO 5218, where the code “2” represents the female gender.

4.3 National Library of Spain

Figure 4 shows the search result for “Austen, Jane” in the catalogue of the National Library of Spain (BNE).

```
leader 01120c1 2200209n 4500
001 001060801094400606
005 20240801122906.0
080 0011209n 011209n 011209n 011209n 011209n
016 ##$aBNE19900162256
024 7$uhttps://viaf.org/viaf/10233412/ $2viaf
024 8$uhttps://data.cervantesvirtual.com/person/471
024 7$uhttps://isni.org/isni/000000012283615X $2isni
024 7$uhttps://www.wikidata.org/wiki/Q1124906 $2wikidata
035 ##$a(Sp)uBNE10X1124906
040 ##$aSp$uBNE $2uBNE $3dc $femhne
046 ##$f1775 $g1817
100 1$A Austen, Jane $d1775-1817
170 ##$aStevenson $uWikidata
372 ##$aFicción gótica $uNovela romántica $uWikidata
374 ##$aNovelistas $uEscritores $uWikidata
375 ##$aFemenino
377 ##$aingles
670 ##$aSentido y sensibilidad, 2014 $pport. (Jane Austen)
670 ##$aOrgullo y prejuicio, 1986 $pport. (Jane Austen)
670 ##$aM. E. 2-4-2014 $uAusten, Jane; 1775-1817
670 ##$aM. E. 2-4-2014 $uJane Austen; n. 16-12-1775 en Steventon-m. 18-7-1817, Winchester; novelista británica contada entre los
935 ##$aX1124906
```

Figure 4. Authority record for the author Jane Austen in the BNE

Field 375 records the author's gender as “Femenino.” Field 100 contains the authorized name with birth and death dates. Field 372 identifies intellectual areas of activity, and 374 specifies professional roles. Field 377 indicates the language of the author's work, while field 670 lists information sources used in the record, such as Wikipedia and the British Library. Field 024 includes external identifiers from VIAF, ISNI, and Wikidata, demonstrating the use of Semantic Web techniques.

These identifiers enhance interoperability, visibility, and data reuse across linked data environments.¹⁸

4.4 National Library of France

Next, Figure 5 presents the authority record in the UNIMARC format in the catalogue of the National Library of France.

```
000 cx a22 45
001 FRBNF11889603X
003 http://catalogue.bnf.fr/ark:/12148/cb118896036
005 20181106
010 .. $a 000000012283635X $2 VIAF $d 20130724
039 .. $o OPL $a 001573266 $z AMA
039 .. $o OPL $a 000419322 $z APP
039 .. $o OPP $a 14871547 $d 20070226
039 .. $o OPP $a 15447424 $d 20121008
100 .. $a 19771109afrey50 ba0
101 .. $a eng
102 .. $a GB
103 .. $a 17751216 18170718
105 .. $a a
106 .. $a 010
120 .. $a a ←
152 .. $c 2
200 .I $7 ba0yba0y $8 fre $9 0 $a Austen $b Jane $f 1775-1817
300 .I. $a Romancière
301 .. $a Steventon (Royaume-Uni) $b Winchester (Royaume-Uni)
801 .. $a FR $b FR-751131015 $c 20181106
810 .. $a Pride and prejudice : a novel / by Jane Austen, 1870 $a Northanger Abbey / Jane Austen ; edited by Barbara M. Benedict and Deirdre Le Faye, 2006
810 .. $a British library - http://catalogue.bl.uk (2011-10-05) $a Oxford English lit., 1991
810 .. $a BN Cat. gén. $a BN Cat. gén. suppl. $a BN Cat. gén. 1960-1969
856 .. $u http://catalogue.bl.uk $e 2011-10-05
```

Figure 5. Authority record in UNIMARC for the author Jane Austen in the BNF

In UNIMARC gender is recorded in field 120, subfield \$a, with the value “a,” which represents female. Field 100 contains the control number, 103 the date of birth, 102 the country code (e.g., GB), and 101 the language code (e.g., a for English). Field 200 shows the authorized name form, and 300 indicates occupation. Field 801 identifies the record’s creator, here the Bibliothèque nationale de France (BnF). Field 856 links to a digital source, such as the British Library catalogue. Fields 810 and 820 list additional consulted sources. Figure 6 presents the same record in INTERMARC format as maintained by the BnF.

¹⁸ According to information provided by the BNE via email, the library conducted a retrospective analysis of its authority records to apply the gender field (375) on a large scale, using the values “Masculino” (male) and “Femenino” (female), without considering other gender identities. Since then, cataloguers have routinely included this information in newly created records. The retrospective process was automated and involved several strategies, such as detecting terms within records that might indicate gender (e.g., *papa*, *rey*, *condesa*), comparing personal names with gendered name lists from the Instituto Nacional de Estadística (INE) of Spain, integrating data from VIAF for foreign authors, and cross-referencing information with Wikidata.

```
000 00778c1 ap22000272 45
001 FRBNF1889603X
008 771109181106gbngf 17751216 18170718 a 010
031 . $a 000000012283635X $2 VIAF $d 20130724
045 . $a a
100 . $w .O.b..... $a Austen $m Jane
600 . $a Romancière
603 . $a Steventon (Royaume-Uni) $b Winchester (Royaume-Uni)
610 . $a Pride and prejudice : a novel / by Jane Austen, 1870 $a Northanger Abbey / Jane Austen ; edited by Barbara M. Benedict and Deirdre Le Faye, 2006
610 . $a British library $u http://catalogue.bl.uk $d 2011-10-05 $a Oxford English lit., 1991
610 . $a BN Cat. gén. $a BN Cat. gén. suppl. $a BN Cat. gén. 1960-1969
917 . $o OPL $a 001573266 $b AMA
917 . $o OPL $a 000419322 $b APP
996 . $o OPP $a 14871547 $d 20070226
996 . $o OPP $a 15447424 $d 20121008
```

Figure 6. Authority record in INTERMARC for the author Jane Austen in the BNF

The authority record for the author Jane Austen, as shown in Figure 6, records gender in field 008, position 17, with the value “f” (female). The presence of field 031, which links the record to VIAF, indicates the application of Semantic Web techniques within the catalogue. This linkage enables interoperability with other databases and reinforces the system’s alignment with Linked Data practices. The inclusion of persistent identifiers and the normalization of access points demonstrate that even within a national format like INTERMARC, the BnF implements mechanisms that support the semantic enrichment of authority records particularly using URIs and connections to internationally recognized external sources.

The analysis of gender representation in authority records across national libraries shows that metadata inclusion reflects a combination of technical decisions, institutional values, ethical concerns, and sociocultural contexts. Libraries adopt varied approaches: some apply only binary values, others introduce broader categories, and some choose to suppress public display of gender data to protect privacy.

These findings echo concerns in the literature that traditional cataloguing frameworks based on rigid data models and predefined Fields struggle to represent complex identity attributes like gender (Adolpho 2019; PCC 2022). Fields such as MARC 375, though designed for this purpose, often reinforce binary or normative classifications that have been increasingly questioned in academic and professional discussions.

This debate is closely linked to Semantic Web principles, Linked Data, and semantic enrichment. These technologies allow authority records to transcend static formats by linking to persistent identifiers (URIs) such as ORCID, VIAF, and Wikidata. This not only enhances interoperability, discoverability, and data accuracy, but also supports individual control over identity data, as advocated by Thompson (2016).

Some national libraries rely on trusted and publicly available sources to determine gender information when it is recorded in authority records. This includes bibliographic references, biographical dictionaries, encyclopedias, and structured linked data sources such as Wikidata and VIAF. These external sources are considered reliable, community-maintained, and aligned with international standards, providing persistent identifiers (URIs) that ensure the accuracy, traceability, and interoperability of the information. When libraries choose to connect their authority records to these datasets, they are not only enhancing metadata quality but also reinforcing ethical practices, since the data is derived from open and verifiable resources. Cataloguing decisions remain guided by ethical frameworks like the IFLA’s Cataloguing Principles (PIC) and the Cataloguing Code of Ethics. Those support interoperability and address concerns about privacy violations, misclassifi-

cation, and the reinforcement of binary gender norms. Ultimately, combining semantic technologies with ethical cataloguing practices leads to more flexible, interconnected, and person-centered authority data.

It requires significant changes in cataloguing workflows, technical infrastructure, and, fundamentally, in the professional mindset regarding authority control. As Possemato (2025) observes, the transformation toward entity-based models and semantic enrichment demands a redefinition of the role of cataloguers not merely as enforcers of data standards, but as mediators of ethical, inclusive, and context-sensitive metadata practices.

The discussion around gender metadata in authority records is far from resolved. It represents an ongoing discussion between the needs for data consistency, ethical responsibility, and the recognition of gender diversity. Moving forward, it is imperative that libraries continue to engage with community-driven initiatives, update their policies to reflect evolving standards, and prioritize approaches that safeguard privacy while promoting equitable representation in the bibliographic universe.

Conclusion

This study shows a range of institutional approaches shaped by metadata standards and cataloguing policies around identity representation. Some institutions adopt inclusive vocabularies that acknowledge gender diversity, while others retain binary models or choose to suppress gender metadata in public catalogues to protect individual privacy. These differences reflect ongoing debates on the role of personal name authority records whether they should record identity attributes or adapt to more fluid, socially constructed understandings of identity.

In this context, semantic technologies can offer viable solutions to enhance metadata practices. By linking authority records to open, persistent identifiers in publicly available and reliable sources like VIAF, Wikidata, and ORCID, libraries can improve accuracy, ensure interoperability, and support personal autonomy. This approach helps mitigate the risks of outdated or incorrect information while aligning with principles of consent, privacy, and inclusivity.

When carefully managed, gender-related metadata can support research, improve data, improve data quality, and acknowledge contributions to cultural and scientific production. However, its presence must always be balanced with privacy concerns and only included when publicly available, verifiable, and consistent with the individual's self-identification.

In light of the findings, it becomes imperative for national libraries and cataloguing institutions to critically assess the ethical dimensions of gender representation in NARs. While the possibility of recording gender information has long existed within the MARC21 authority format most notably through fields such as 375 (Gender), what is now being proposed is not merely the inclusion of this data, but a critical revision of how it is collected, structured, and exposed. This involves moving away from inferential, binary, and static models toward approaches grounded in consent, self-identification, and ongoing revision. Beyond adopting inclusive descriptive standards and metadata schemas, institutions must reevaluate cataloguing codes and controlled vocabularies to better reflect the complexity and fluidity of gender identities. This includes implementing mechanisms for user consent, enabling updates to authority records based on individual agency,

and developing ethical policies that prevent the reinforcement of normative classifications or the inadvertent outing of individuals. By embedding ethical review and participatory frameworks into authority control workflows, libraries can advance toward a more inclusive, respectful, and socially responsible bibliographic infrastructure.

References*

- Adolpho, Kalani. 2019. "Who Asked You? Consent, Self-Determination, and the Report of the PCC Ad Hoc Task Group on Gender in Name Authority Records." In *Ethical Questions in Name Authority Control*, edited by Jane Sandberg, 111–31. Sacramento, CA: Library Juice Press.
- Berners-Lee, Tim. 2006. "Linked Data." *Design Issues*, July 27. <https://www.w3.org/DesignIssues/LinkedData.html>.
- Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. "The Semantic Web." *Scientific American*. https://www-sop.inria.fr/acacia/cours/essi2006/Scientific%20American_%20Feature%20Article%20The%20Semantic%20Web_%20May%202001.pdf.
- Billey, Amber, Emily Drabinski, and K. R. Roberto. 2014. "What's Gender Got to Do with It? A Critique of RDA 9.7." *Cataloging & Classification Quarterly* 52 (4): 412–21. <https://doi.org/10.1080/01639374.2014.882465>.
- Burns, Jasmine, Michelle Cronquist, Jackson Huang, Devon Murphy, K. J. Rawson, Beck Schaefer, Jamie Simons, Brian M. Watson, and Adrian Williams. 2022. "Metadata best practices for trans and gender diverse resources." *Librarian Publications & Presentations* 52. <https://digitalcommons.osgoode.yorku.ca/librarians/52>.
- Butler, Judith. 1990. *Gender Trouble: Feminism and the Subversion of Identity*. New York, NY: Routledge.
- Carbajal, Itza A. 2020. "Review of Ethical Questions in Name Authority Control." *Journal of Contemporary Archival Studies* 7: Article 7. <https://elischolar.library.yale.edu/jcas/vol7/iss1/7>.
- Gil, Antonio Carlos. 2008. *Métodos e Técnicas de Pesquisa*. São Paulo: Atlas.
- IFLA (International Federation of Library Associations and Institutions). 2016. "Statement of International Cataloguing Principles." https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/icp/icp_2016-pt.pdf.
- Isaac, Antoine, Hugo Manguinhas, Juliane Stiller, and Valentine Charles. 2015. *Report on Enrichment and Evaluation*. The Hague, Netherlands: Europeana Task Force on Enrichment and Evaluation. http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/Enrichment_Evaluation/FinalReport_EnrichmentEvaluation_102015.pdf.
- Library of Congress. 2018. "MARC 21 Format for Authority Data." <https://www.loc.gov/marc/authority/ecadhome.html>.
- Library of Congress. 2022. "375 (Gender) R." <https://www.loc.gov/marc/authority/ad375.html>.
- Lindsey, Nerissa, Kuriger Suiter Greta, Hanselman Kurt. 2022. Ethical Considerations of Including Gender Information in Open Knowledge Platforms." *KULA: Knowledge Creation, Dissemination, and Preservation Studies* 6 (3): 1–15. doi: [10.18357/kula.228](https://doi.org/10.18357/kula.228).
- Manguinhas, Hugo, Nuno Freire, Antoine Isaac, Juliane Stiller, Valentine Charles, Aitor Soroa, Rainer Simon, and Vladimir Alexiev. 2016. "Exploring Comparative Evaluation of Semantic En-

* Last date of websites' consultation: January 9, 2026.

richment Tools for Cultural Heritage Metadata.” In *Research and Advanced Technology for Digital Libraries: 20th International Conference on Theory and Practice of Digital Libraries, TPDL 2016, Hannover, September 5–9, Proceedings 20*, 266–78. Springer International Publishing.

Marcondes, Carlos Henrique. 2001 “Representação e economia da informação.” *Ciência da Informação* 30 (1): 61–70. <http://www.scielo.br/pdf/ci/v30n1/a08v30n1>.

Martin, Jennifer M. 2021. “Records, Responsibility, and Power: An Overview of Cataloging Ethics.” *Cataloging & Classification Quarterly* 59 (2–3): 281–304. <https://doi.org/10.1080/01639374.2020.1871458>.

National Information Standards Organization. 2022. “Gender Identification and Cataloging Practice.” *NISO I/O*, May. <https://www.niso.org/niso-io/2022/05/gender-identification-and-cataloging-practice>.

Oliver, Chris. 2021. *Introducing RDA: A Guide to the Basics after 3R*. 2nd ed. Chicago: ALA Editions.

Olson, Hope A. 2002. *The Power to Name: Locating the Limits of Subject Representation in Libraries*. Dordrecht: Kluwer Academic Publishers.

PCC Ad Hoc Task Group on Gender in Name Authority Records. 2022. *Revised Report on Recording Gender in Personal Name Authority Records*. <https://www.loc.gov/aba/pcc/documents/gender-in-NARs-revised-report.pdf>.

Pomerantz, Jeffrey. 2015. *Metadata*. Massachusetts: The MIT Press.

Possemato, Tiziana. 2025. “Recording Gender in the Person Entity: An Ongoing Discussion.” *JLIS.it* 16 (1): 1–12. <https://doi.org/10.36253/jlis.it-630>.

Prodanov, Cleber Cristiano, and Ernani Cesar de Freitas. 2013. *Metodologia do Trabalho Científico: Métodos e Técnicas da Pesquisa e do Trabalho Acadêmico*. Novo Hamburgo: Feevale. <https://www.feevale.br/institucional/editora-feevale/metodologia-do-trabalhocientifico--2-edicao>.

Scott, Joan Wallach. 1995. “Gender: A Useful Category of Historical Analysis.” Translated from the French version by Guacira Lopes Louro; revised by Tomaz Tadeu da Silva based on the original English text. *Educação & Realidade* 20 (2): 71–99. <https://lume.ufrgs.br/bitstream/handle/10183/257862/000037108.pdf>.

Seikel, Michele, and Thomas Steele. 2011. “How MARC Has Changed: The History of the Format and Its Forthcoming Relationship to RDA.” *Technical Services Quarterly* 28 (3): 322–334. <https://doi.org/10.1080/07317131.2011.574519>.

Smith-Yoshimura, Karen. 2018. “The Rise of Wikidata as a Linked Data Source.” *Hanging Together*. <https://hangingtogether.org/the-rise-of-wikidata-as-a-linked-data-source/>.

Taylor, Arlene. G., and Daniel N. Joudrey. 2009. *The Organization of Information*. Westport: Libraries Unlimited.

Thompson, Kelly Jean. 2016. “More than a name: A content analysis of name authority records for authors who self-identify as trans.” *Library Resources & Technical Services* 60 (3): 140–55.

Tillett, Barbara. 2004. "Authority control: State of the art and new perspectives." *Cataloging & Classification Quarterly* 38(3–4): 23–41. https://doi.org/10.1300/J104v38n03_04.

United Nations. 2015. "Goal 5: Achieve gender equality and empower all women and girls" Sustainable Development Goals." United Nations. <https://www.un.org/sustainabledevelopment/gender-equality/>.

Wainer, Luiza. 2019. "Questão de gênero: considerações sobre a inclusão de gênero em registros de autoridade." In *Anais do I Encontro de RDA no Brasil*. Florianópolis: Editora da UDESC. <https://rdanobrasil.org/wp-content/uploads/2019/12/2019-questao-de-genero-em-registros-de-autoridade-texto.pdf>.

Zeng, Marcia Lei. 2019. "Semantic Enrichment for Enhancing LAM Data and Supporting Digital Humanities. Review Article." *Profesional de la Información* 28 (1). <https://doi.org/10.3145/epi.2019.ene.03>.

Hipátia: a model for the digital preservation of records and information?

Angelica Alves da Cunha Marques^(a)

Karoline dos Santos Rodrigues^(b)

Denise Oliveira de Araújo^(c)

Tauane Fonseca Esteves^(d)

Marilete da Silva Pereira^(e)

Alexandre Faria de Oliveira^(f)

Tiago Emmanuel Nunes Braga^(g)

a) University of Brasilia, <https://orcid.org/0000-0003-4642-5912>

b) Brazilian Institute of Information in Science and Technology, <https://orcid.org/0009-0000-3619-1070>

c) Brazilian Institute of Information in Science and Technology, <https://orcid.org/0000-0003-4033-7882>

d) Brazilian Institute of Information in Science and Technology, <https://orcid.org/0009-0008-2549-3418>

e) Brazilian Institute of Information in Science and Technology, <https://orcid.org/0000-0003-3382-4084>

f) Brazilian Institute of Information in Science and Technology, <https://orcid.org/0000-0003-0470-4972>

g) Brazilian Institute of Information in Science and Technology, <https://orcid.org/0000-0001-6332-7965>

Contact: Angelica Alves da Cunha Marques, angelicacunha@unb.br;

Karoline dos Santos Rodrigues, karolinerodrigues@ibict.br; Denise Oliveira de Araújo, denisearaujo@ibict.br;

Tauane Fonseca Esteves, tauaneesteves@ibict.br; Marilete da Silva Pereira, mariletepereira@ibict.br;

Alexandre Faria de Oliveira, alexandreoliveira@ibict.br; Tiago Emmanuel Nunes Braga, tiagobraga@ibict.br

Received: 11 December 2025; **Accepted:** 26 March 2026; **First Published:** 15 May 2026

ABSTRACT

This study investigates whether Hipátia can be characterised as a digital preservation model for records and information. Developed by the Brazilian Institute of Information in Science and Technology (Ibict), Hipátia functions as an infrastructure for integration and automation among information systems, aimed at implementing Trusted Digital Archival Repositories (TDARs). Motivated by the lack of conceptual consensus surrounding the term “model” in scientific literature, the article offers a theoretical and epistemological examination of Hipátia’s relevance to Information Science and Archival Science. It employs a qualitative, exploratory, and descriptive approach grounded in a literature review. Works from Philosophy and Information Science addressing the notion of model were analysed, alongside technical and scientific studies on Hipátia, especially its configuration as a TDAR and its alignment with the Open Archival Information System (OAIS) model. Findings suggest that, although Hipátia does not fully conform to traditional definitions, it can be framed as a Kuhnian exemplar, an analogical model (Hesse), an operational representation (Abrantes), an applied method (Dutra), and a logical structure (Silva). Its databus (BarraPres) and capacity for integrating diverse systems highlight its flexibility, interoperability, and diagnostic potential. The study concludes that Hipátia operates not only as a technical tool but as a multifaceted model.

KEYWORDS

Hipátia; Paradigm; Model; Digital Preservation.

1. Introduction

In scientific literature, there is no consensus on the polysemic notion of “model” (Dutra 2021), even though it is used by all scientific disciplines. Philosophy dictionaries contain a variety of definitions, and one definition related to the idea of paradigm stands out (Japiassú and Marcondes 1996, 184), in which this notion is presented as follows: “Paradigm, ideal form. An object that serves as a parameter for the construction or creation of others”.

These dictionaries also define metaphor as a figure of speech in which a symbolic comparison is made between elements using a word with a transposed meaning. According to Japiassú and Marcondes, a metaphor is a “word used with an altered meaning”, for instance, expressions like “reinventing the wheel” or “push the right buttons” (Japiassú and Marcondes 2001, 128). Although it is not always explored explicitly, the metaphor shares certain cognitive mechanisms of representation with the model and will be analysed in this article based on the distinction proposed by Silva (2010), who contrasts it with the model in terms of its stylistic and interpretative function.

The relationship between the definition of model and the notion of “paradigm” proposed by Thomas Kuhn¹ – a North American physicist who became a philosopher and historian of science (Schulz 2023) – is inevitable in his work *The Structure of Scientific Revolutions*, published in 1962. The study of contemporary science based on scientific ideas and their truths considers “[...] a succession of tradition-bound periods punctuated by non-cumulative breaks” (Kuhn 1970, 208), in other words, the paradigms. From this perspective, Kuhn starts from the history of science, moves on to generalisations about the psychosocial conditions of its development, and arrives at a new proposal for epistemology (Oliva 1994), whose main definitions – scientific community, paradigms, normal science and scientific revolutions – refer to the parameters of development of the Natural Sciences (Physics, Biology, Astronomy, etc.), that is, to what he calls “Normal Science” (Oliva 1994, 81).

Despite being heavily criticised, his work became a reference not only for the Natural Sciences, but also for the Social Sciences (Oliveira 1997), such as Information Science and Archival Science, in which several works take Kuhn’s notions as theoretical references, such as Eugênio, França, and Perez (1996), Nehmy et al. (1996), Sayão (2001), Renault (2007) and Marques (2011).

In a similar exercise of searching for definitions of “paradigm” in reference works, this time in a dictionary limited to the field of Information Science, the notion is understood as a representation for large-scale reproduction or execution (Cunha and Cavalcanti 2008). The definitions listed by Cunha and Cavalcanti address the paradigm in the context of linguistics, positioning it as an element that has at least one common attribute or as a “set of units that are likely to appear in the same context, and are therefore interchangeable and mutually exclusive” (Cunha and Cavalcanti 2008, 271). The authors also fail to mention what is perhaps the most popular meaning of the term “paradigm”, namely the understanding theorised by Thomas Kuhn (Cunha and Cavalcanti 2008). Kuhn’s work is directly and indirectly taken up by other scholars of philosophy, including his con-

¹ Thomas Kuhn was born in Cincinnati in 1922. He read Physics at Harvard and developed an interest in the History of Philosophy through courses in History and Philosophy taken during his undergraduate and postgraduate studies, respectively. In an interview given to Baltas, Gavroglu and Kindi (Athens, 19–21 October 1995), Kuhn remarked: “I wanted to teach myself enough history of science to establish myself there in order to do the philosophy” (Baltas, Gavroglu, and Kindi 2000, 276).

temporary Mary Hesse² (1924–2016), who understands a model as an “imperfect copy” (Hesse 1966), that is, a representation of a reality or phenomenon that seeks to explain or predict its functioning. In the Brazilian context, Abrantes (2020) stands out with his study, which, from a philosophical-epistemological perspective, understands a model as a set of physical entities or linguistic propositions. In turn, his fellow countryman, Dutra (2021) presents different views of models in different areas of knowledge: abstract constructions that serve as cognitive tools to represent and understand phenomena in both the natural sciences and the social sciences.

For a philosophical approach to Information Science, we find in Silva’s work (2010) a more procedural approach in which the model, a map representing the characteristics of its territory, serves for modelling, linking to behaviours that refer to the user’s informational behaviour in the search for information. Silva revisits several studies in philosophy to delimit the definition of “model”, unfolded into “modelling”, and considers the “ongoing paradigmatic transition” (Silva 2010, 322) within the scope of Information Science.

Given the above, this article aims to map definitions of the notion of “model” to support the characterisation of Hipátia, proposed by the Brazilian Institute of Information in Science and Technology (Ibict), as a model for the digital preservation of records and information. To achieve the goal, we chose a qualitative, exploratory, and descriptive approach. It is developed through bibliographic research, which maps authors of Philosophy and Information Science around theoretical references on the aforementioned notion, in addition to studies of Information Technology (IT) and Archival Science on the definitions and characteristics of Hipátia.

The text is divided into four parts: this introduction; section 2, which examines the theoretical references of philosophy and Information Science on the notion of model; section 3, which presents limitations and possibilities of Hipátia as a model; and the final remarks.

2. Literature review: theoretical references surrounding the notion of “model”

As mentioned in the introduction, there is no consensus on the notion of “model” in scientific literature, although it is widely and frequently used across all disciplines. Recognising the limitations of a scientific article, we chose to read Kuhn’s work (1970), considered a fundamental theoretical reference for understanding models based on paradigms.

² “Mary Brenda Hesse, born 15 October 1924, studied first at Imperial College then at University College London, receiving her doctorate for a thesis on electron microscopy in 1948. She lectured on mathematics at Royal Holloway College from 1947 to 1951 and at the University of Leeds from 1951 to 1955, then from 1955 to 1959 on history and philosophy of science at University College. In 1960 she was appointed to a lectureship in history and philosophy of science at Cambridge, from 1968 to a readership and from 1975 to a professorship. [...] Mary’s important philosophical works were marked from the outset (in her pioneering *Forces and Fields*, 1961) by close engagement with the history of the practices and theories of the sciences. Areas in which her studies had major impact include the roles of metaphor and analogy in scientific investigation (*Models and Analogies in Science*, 1963) and the relations and interactions of observation, experiment and theory in the sciences (*The Structure of Scientific Inference*, 1974)”. <https://www.bps.cam.ac.uk/news-events/news-archive/mary-hesse>.

³ The direct citations of the following texts were translated from the original portuguese into english: Abrantes (2020), Braga (2022a; 2022b), Braga, Holanda, and Canelhas (2022), Dutra (2021), Japiassú and Marcondes (2001), Oliveira (1997), Oliveira (2022), Oliveira, Shintaku, and Silva (2022), Renault (2007), Saad and Brito (2024), Sayão (2001), Shintaku, Braga, and Oliveira (2021), Shintaku and Oliveira (2022), Shintaku et al. (2023), Silva, Cardoso, and Brito (2005), and Silva (2010).

2.1. Thomas Kuhn: paradigms, models, and exemplars

Inspired by the presentation of Physics as a science for non-scientists and concerned with the philosophical problems of “historically oriented science”, Kuhn (1970, 15) formulates some fundamental concepts for understanding what he considers to be the “[...] history of scientific ideas”, in a predominantly internalist approach, but one that does not ignore the relevance of factors extrinsic to science. With the aim of “knowing what is true”, that is, “to know what it is to be true” (Baltas, Kostas, and Kindi 2000, 278), the scholar is concerned with the use of meanings associated with changing values and argues that “it’s not only the ideas that change, it’s the structure of the disciplines that are working on them” (290).

Far from traditional epistemological tradition,⁴ Kuhn seeks to understand the construction of scientificity in its historical aspect, in an approach he himself considers conservative, although he does not intend to label what is “science” or “pseudoscience” (Baltas, Kostas, and Kindi 2000).

Although this work falls within the scope of Applied Social Sciences,⁵ it is considered appropriate, as in many studies in this field, to revisit and use as a reference the concepts proposed by Kuhn regarding scientific development and the notion of “paradigm”, which, by imposing practices of observation and theorisation that formulate a tacit worldview, gives rise to the tradition of investigation (normal science), which represents a type of research that differentiates science from non-science (Oliva 1994).

In his classic work, Kuhn (2000) later clarifies some aspects of his initial formulation. According to the author:

- a) the notion of normal science was linked to solving puzzles, by perfecting and extending the correspondence between theory and practice in the development of the field (Baltas, Gravoglu, and Kindi 2000; Kuhn 2000), that is, “[...] is what produces the bricks that scientific research is forever adding to the growing stockpile of scientific knowledge” (Kuhn 2000, 13);
- b) on the other hand, scientific revolutions “[...] involve discoveries that cannot be accommodated within the concepts in use before they were made. In order to make or to assimilate such a discovery one must alter the way one thinks about and describes some range of natural phenomena” (Kuhn 2000, 14-5);
- c) finally, the traditional idea of paradigms was that of models, “[...] particularly grammatical models of the right way to do things” (Baltas, Gravoglu, and Kindi 2000, 298), as well as the notion of exemplars within the structure and that of “[...] hermeneutic basis for the science of a particular period” (Kuhn 2000, 221).

The author turns to the functioning of the scientific community, highlighting the techniques of persuasive argumentation that permeate it. For him, the constitution of a field is linked to the definition of what will be considered an authentic problem or an adequate solution within it (Oliva

⁴ Oliva (1994) recalls that Kuhn was influenced by French epistemology, unlike most Anglo-American philosophers, who focused on the study of logical-empirical reasons, forgetting the factors extrinsic to “scientific reason” that are indispensable for understanding the functioning of scientific communities.

⁵ Considering Information Science and Archival Science in the classification of areas of knowledge of the National Council for Scientific and Technological Development (CNPq), currently in force. Available at: <https://lattes.cnpq.br/documents/11871/24930/TabeladeAreasdoConhecimento.pdf>.

1994). Kuhn conceives of a culture in which the concept of the object to be exemplified/studied is available, that is, transmitted from one generation to another, especially through manuals (Kuhn 2000). He seeks to describe, in the field of Natural Sciences, “[...] the way a tradition worked in terms of consensus” (Baltas, Gravoglu, and Kindi 2000, 299), based on the principle that “[...] the natural sciences of any period are grounded in a set of concepts that the current generation of practitioners inherit from their immediate predecessors” (Kuhn 2000, 221).

Thus, in the afterword to *The Structure of Scientific Revolutions*, published in 1962, Kuhn (1970) redefines the notion of paradigm, emphasising the concept of exemplars as sets of problems and standard solutions that embody the consensus within the scientific community. Exemplars are used to guide the practice of scientists during a period of normal science, thus becoming fundamental in the training of scientists, since they are transmitted through manuals and problem-solving. The expectation is that, through modelling based on these standards, scientists will be able to solve new problems throughout their scientific practice.

Also in the afterword, the scholar replaces the term “theory” with “disciplinary matrix”: “[...] ‘disciplinary’ because it refers to the common possession of the practitioners of a particular discipline; ‘matrix’ because it is composed of ordered elements of various sorts, each requiring further specification” (Kuhn 1970, 182).

Considering Kuhn’s relevant reflections and contributions on scientific development, we revisit some of his main ideas, while recognising their limitations when recontextualised in the social sciences, as Nehmy et al. (1996), Eugênio, França, and Perez (1996), Sayão (2001) and Renault (2007) in the study of Information Science. Although Kuhn defends the existence of only one paradigm in a given period, it is possible to question the coexistence of paradigms and scientific revolutions in the constitution of the scientific community, as discussed by Oliveira (1997) in the field of Anthropology. Given his paradigms and schools, this author interprets Kuhn and presents his definition of “disciplinary matrix”, adapting it to the humanities: “[...] a disciplinary matrix is the systematic articulation of a set of paradigms, the condition of their coexistence in time, with each and every one remaining active and relatively efficient” (Oliveira 1997, 15), a notion that can be appropriated in Information Science.

2.2. Mary Hesse: Analogies and Models in Science and Theology

Hesse defines “model” as a fundamental concept in science, used to represent and explain phenomena through analogies and metaphors. In her approach, models are not merely auxiliary tools, but essential components in the structure of scientific theories. The author argues that scientific models work through analogies between known and unknown systems, allowing inferences about the behaviour of complex systems based on their structural similarities with other systems.

Her definition of model, in the book *Models and Analogies in Science* (1966), appears in a more systematic way at the beginning of the chapter “The Function of Models: A Dialogue”, in a dialogue between two fictional characters: “Campbellian” and “Duhemist”, when the former proposes a distinction between types of analogy (positive, negative, and neutral).

Let us agree that, when we speak of a model in its primary sense (in this discussion let us call it model₁), we are not speaking of another object which can, as it were, be built or imagined alongside

the phenomena we are investigating. The model₁ is the imperfect copy (the billiard balls) minus the known negative analogy, so that we are only considering the known positive analogy, and the (probably open) class of properties about which it is not yet known whether they are positive or negative analogies (Hesse 1966, 9).

By this definition, the model is understood to be a system of structured analogies, involving:

- positive analogies (known valid similarities between the model and the real system);
- negative analogies (known differences);
- neutral analogies (aspects that are not yet known to be similar or not, and in which the heuristic potential of the model lies).

In Hesse's view (1966), it is a representation of a reality or phenomenon that seeks to explain or predict its functioning. Models can be physical, mathematical, or logical, and are used to understand the essential structures of a system, serving as instruments that allow analogies between different areas of knowledge. The author's main ideas about models include:

- Representation and explanation: models are used to represent structures and processes, allowing for a better understanding of the phenomena studied.
- Analogical nature: the correspondence between a model and reality is often partial and based on structural similarities.
- Flexibility and evolution: models are not static; they advance as new information, and experiments modify our understanding of phenomena.
- Limitations and applicability: no model is completely accurate, but it provides useful approximations that must be evaluated in terms of their effectiveness in predicting and explaining data.

2.3. Paulo Abrantes: method and Science

Abrantes⁶ (2020) addresses the ambiguity and conceptual diversity surrounding the notion of a model, both from a philosophical and epistemological perspective, considering logical empiricism. For him, models can be understood as sets of physical entities or as sets of linguistic propositions. In addition, they are often used as representations of real systems and can be symbolic (mathematical equations), iconic (drawings, maps), or material.

Kuhn's ideas about normal science are related to the models mentioned by Abrantes (2020), who considers paradigms as shared models and as exemplars in normal science, which establish guidelines that, in times of crisis, can be questioned and replaced.

In Abrantes' view (2020), scientific models are defined as fundamental tools for applying theories, as well as for explaining, predicting, and controlling phenomena. Thus, they play a crucial role in testing theories and simulating phenomena, allowing for a deeper understanding of the behaviour of

⁶ "Paulo C. C. Abrantes holds a degree in Physics from the University of Brasília (1973), a master's degree in Philosophy from the University of Paris X (1978) and a doctorate in Philosophy from the University of Paris I (1985). He is a retired full professor at the University of Brasília, having worked in the Department of Philosophy and the Institute of Biological Sciences. His research focuses mainly on the areas of Philosophy of Science, Philosophy of Biology, History of Science, and Philosophy of Mind". <https://www.posfil.unb.br/index.php/2-publicacoes/111-prof-dr-paulo-cesar-coelho-abrantes>.

complex systems. Furthermore, models can be differentiated from theories because they are idealised representations of particular physical systems, while theories are more comprehensive and general. In short, his work emphasises that models are fundamental constructs in science, but they have limitations that influence their applicability and interpretation, whether in their choice conditioned by the mathematical tools available, which can lead to forced simplifications, or by mental habits that can influence the design of models, favouring those that have analogies with familiar systems. Considering the ambiguity of the term “model”, Abrantes states: “The term ‘model’ is ambiguous, being used by both scientists and philosophers with a plurality of meanings. An important methodological task is to categorise such models, show the contexts in which they are used, and for what purposes” (Abrantes 2020, 215). He then categorises them into four types, as shown in Table 1.

Table 1. Model types based on Abrantes (2020)

Model types	Definition
Semantic (type 1)	“A model, in this sense of the term, is an interpretation of a calculus, that is, a structure of objects, properties, and relations that satisfy the sentences of a calculus, making them true propositions about a particular world” (p. 216).
Analogical (type 2)	“Familiar, widely accepted and corroborated theories are often the source of analogical models for generating other theories applicable to new domains of phenomena. In this second sense, modelling is a relationship between two representations, for example, two theories” (p. 216)
As representations of real systems (type 3)	“To be a representation of a system, the model must bear some similarity to the system. [...] A model does not represent all the entities that make up a system, with all their properties and interrelationships. If that were the case, the model would be a duplication of the system and would not fulfil its function. The model in this sense is, therefore, an idealised representation insofar as it abstracts characteristics of the system that are considered irrelevant” (p. 217-218).
Abstract (type 4)	“Abstract structures can be purely syntactic (formal, uninterpreted symbolic structures) or interpreted, such as mathematical structures” (p. 219).
Concrete models of real systems (type 5)	“Concrete models and the real system being modelled can be seen as different materialisations, or realisations, of a mathematical structure (and here we return to type 4 models). This type of model is often used to simulate the behaviour of a system. [...] Integrating subtypes 4 and 5 into the more general characterisation I made of the third type of model, modelling is here a relationship between a representation (symbolic, iconic or material) and a real system” (p. 220-221).

2.3. Luiz Henrique Dutra: models in Natural and Social Sciences

Dutra⁷ (2021) discusses the importance of scientific models and their relationship with reality. He presents them as abstract constructions that serve as cognitive tools for representing and under-

⁷ Luiz Henrique de Araujo Dutra holds a bachelor’s degree in Philosophy from the University of São Paulo (USP) and a master’s and doctorate in Logic and Philosophy of Science from the State University of Campinas (UNICAMP). He is a professor at the Federal University of Santa Catarina (UFSC) and teaches in the postgraduate programmes in Philosophy at UFSC and UnB. <http://lattes.cnpq.br/7589426150000881>.

standing phenomena in both the Natural Sciences and the Social Sciences. He employs the notion of the “nomological machine”, a concept developed by Nancy Cartwright (1999, quoted in Dutra 2021, 130), to describe scientific models as structures that exhibit mechanical laws or recurring patterns in the world. In this sense, nomological models are not exclusive to the Natural Sciences and are present in the Humanities and everyday knowledge.

Therefore, among other differences always pointed out between the natural sciences and the cultural sciences is irrelevant, namely, the lack of theories in the cultural sciences. For if there is a theory, the latter is just one more source of models. However, if there is no theory, the models will emerge anyway. And it is with them that we will always work. This modelling activity is the basic and unified methodology of all scientific disciplines, in any field, as well as the methodology of everyday life. Models do in the domain of method what emergence does in the domain of ontology (Dutra 2021, 133).

The scholar emphasises that modelling is an essential activity for human cognition, going beyond the sciences and permeating the way we interact with the world.

2.4. Armando Malheiro da Silva and other Information Science studies on models and modelling

The notion of model is approached from different perspectives in the search for models and modelling for Information Science, undertaken by Silva.⁸ The scholar revisits the origins of the word “model” to define it:

The etymological root of the word is *modus*, which means measure, and, interestingly, this meaning remains alive in the scientific field: a model is a physical, mathematical or logical system that represents the essential structures of a reality and is capable, at its level, of dynamically explaining or reproducing the functioning of these same structures (Silva 2010, 299).

Then, the author explains the difference between model and metaphor:

In any of the categories, it is worth noting that the model is a stylisation, but it should be pointed out that a stylised version of reality is not a metaphor for reality, quite the contrary. [...] Models and metaphors are operations with a common root in thought, but they function in opposite ways – the former empty images to bring them closer to univocity, while the latter fill images with a multiplicity of connections to increase their ambiguity (309)

On this point, Mary Hesse’s (1966) conception of metaphor offers an important counterpoint. For Hesse, metaphor should not be understood merely as a linguistic ornament, but as a fundamental cognitive resource in scientific development. She argues that metaphors operate through the interaction between two domains—the primary domain, which is sought to be explained, and

⁸ He holds a degree in Philosophy from the Faculty of Philosophy and Social Sciences of the Portuguese Catholic University, Braga; in History from the Faculty of Arts of the University of Porto; and in Library and Archival Studies from the Faculty of Arts of the University of Coimbra. He holds a PhD in Contemporary History from the University of Minho. He is currently a professor in the Information Science Degree Programme at the University of Porto.

the secondary domain, the source of the metaphor—allowing the creation of new relationships of meaning that transform the understanding of both domains. Rather than solely generating ambiguity, as Silva (2010) suggests, Hesse highlights the heuristic and explanatory function of metaphor, especially when introducing new theoretical vocabularies.

Although Silva (2010) sees metaphor as a producer of ambiguity, Hesse (1966) demonstrates that its epistemic function can be precisely the opposite: to contribute to the formulation of theoretical models by allowing the conceptual redescription of phenomena. By acting as a preliminary stage of modelling, metaphor plays an essential role in the construction of knowledge, not as an obstacle, but as a path to interpretative innovation. As Hesse states: “the deductive model of explanation should be modified and supplemented by a view of theoretical explanation as metaphoric redescription of the domain of the explanandum” (Hesse 1966, 157).

This reinterpretation reinforces the idea that metaphors and models are not mutually exclusive categories. On the contrary, they operate at complementary levels of scientific production: while the model seeks to logically represent reality, the metaphor contributes to the emergence of new modes of representation. In this sense, metaphor can be understood as a resource that enables theoretical modelling itself, especially in interdisciplinary fields such as Information Science.

According to Silva (2010), a model can be understood as a representation of reality, whether in physical, mathematical or logical form, with a view to explaining or reproducing the essential structures of that reality, either totally (covering a phenomenon in its entirety) or partially (focusing on a specific aspect).

The same author states that, in the Social Sciences, models are used to simplify complex systems and make them amenable to quantitative analysis, as in the study of information behaviour and infometrics. In Information Science, on the other hand, information management models, for example, are often more logical or conceptual than mathematical, being used to solve problems and optimise results in information retrieval.

Within the scope of this discipline and his proposal to distinguish between models and modelling, Silva draws on Paul Otlet’s *Traité de Documentation* (1934) to indicate models and modelling from a cumulative or fragmentary perspective, which, according to him, “[...] consists of not being clear or visible in terms of their theoretical basis” (Silva 2010, 323). In this quest, he draws on the work of González Teruel (2005) to understand models of information behaviour within the scope of information seeking (needs and uses), in which, according to him, “These are examples of modelling, based on the empirical principle that theorisation does not precede the model, but rather it is the model that, by establishing patterns and enabling generalisations, becomes decisive in the development of theory” (Silva 2010, 325).

When presenting the eLit.pt Model and research on information literacy, he refers to the theoretical and methodological tools of Information Science, taking into account information production, organisation and representation, and information behaviour within the scope of information management, “[...] which focuses on problems and solutions, from production to use, covering all stages of the information-communication cycle” (341).

These operational concepts play a decisive role in the development of an Information Science model, in the face of problems, cases, and situations related to information behaviour. According to the author, eLit.pt falls into the category of models that focus on people’s behaviour in relation to the information they seek and communicate for use in specific situations, contexts, and environ-

ments. In this sense, it is a model that represents and diagnoses, as well as guides more practical interventions.

Still within the scope of Information Science, Renault⁹ (2007) recognises the robustness of the paradigm concept proposed by Thomas Kuhn. However, given the particularities of Information Science, he opts for a hybrid approach that starts from epistemology and moves between the humanities and social sciences, based on Ivan Domingues.¹⁰ In this sense, the author points out that the notion of paradigm was inaugurated by Plato, which influences the understanding we have today. Nevertheless, Renault (2007) considers that both the etymology of the term “paradigm”, which “[...] comes from the Greek *paradigma* and means model [...]”, and the Platonic heritage, interpreted imprecisely, which suggests the concept as something “exemplary” or “model”, corroborate the erroneous attribution of a synonymic relationship between “paradigm” and “model” (53).

In summary, Renault (2007) presents the “paradigm” as a lens, that is, a worldview or way of seeing a phenomenon, directly associated with the notion of “theory” and capable of encompassing numerous models. The “model”, in turn, would be a representation of reality or of a system and could be explained by different theories, having a concrete, applied, and operational nature, and therefore being at the methodological level. From this perspective, the model is seen as an integrated part of a paradigm (or theoretical system), simply operating without necessarily promoting structural ruptures, as occurs with paradigms in the sense recognised by Kuhn (1970).

The approach presented by Renault (2007) is close to the concepts of Abrantes (2020) and Dutra (2021) in relating the “model” to the notion of method. Abrantes (2020), in particular, by viewing models as a tool for applying theories, can be drawn closer to Renault’s (2007) epistemological approach.

Sayão¹¹ (2001) discusses models as methodological tools and instruments of abstraction, essential for understanding and representing reality. For the author, a model fulfils descriptive, explanatory, and, in some cases, predictive functions by presenting a heuristic dimension, revealing previously unnoticed aspects of the observed reality. As a representation, a model always implies a reduction

⁹ Graduated in Library Science, Master’s, and Doctorate in Information Science from the Federal University of Minas Gerais (UFMG). “Has experience in the field of Information Science, with an emphasis on Information Theory. He works in the University Libraries segment, participating in activities involving projects, management, and information technology. He is interested in topics related to the Epistemology of Information Science, relationships between Archives, Museums, and Libraries, Reading Mediation, Memory, Culture and Information, and Information and Communication Technologies” (Renault 2023).

¹⁰ See: (Domingues 2004)

¹¹ He holds a degree in Physics and a master’s and doctorate in Information Science from the Federal University of Rio de Janeiro (UFRJ). “He has worked at the National Nuclear Energy Commission (CNEN) since 1980. [...] He was an advisor to CONARQ - National Archives Council and the Technical-Scientific Committee of IBICT. He is currently the leader of research groups in Nuclear Knowledge Management; coordinator of the technological development project Infrastructure to Support the Management and Preservation of Nuclear Knowledge; alternate representative of Brazil at INIS (IAEA/UN); member of CONARQ’s Technical Chamber of Electronic Documents; representative of CNEN at OGP/commitment 4 (Open Government Partnership); permanent professor in the Postgraduate Programme in Information Science at IBICT/UFRJ and collaborating professor in the Postgraduate Programme in Library Science at UNIRIO - Federal University of the State of Rio de Janeiro and the Postgraduate Programme in Memory and Collections at the Casa de Rui Barbosa Foundation; representative of the nuclear area at GO FAIR BRASIL; member of the Steering Committee of FAIR Digital Object” (Sayão 2025).

in complexity, prioritising relevant elements according to the modelling objective.

Modelling, in this sense, consists of the process of constructing these models and entails specific methodological and conceptual choices. According to Sayão (2001), information modelling in Information Science consists of identifying the components of models, their modes of expression, and underlying paradigms, functioning as a useful mechanism for capturing and communicating knowledge about the search, organisation, and use of information.

Furthermore, according to Burt and Kinnucan (1990, quoted in Sayão 2001), a continuous configuration for information models is presented, composed of two extremes: on one side, human beings with their internal, subjective, and idiosyncratic reality; on the other, the information system, with its technical and operational structure. Between these extremes lies the field of representations, an intermediate space in which models seek to build bridges and interpret the relationships between individuals and systems.

In this continuum, models are classified as follows:¹²

Models that closely represent human users and their mental processes in relation to the system are called cognitive models, whereas models that focus on the system itself and attempt to describe its inner workings are better known as data models. In the middle of the continuum are geometrical models that interpret users, the system, and the interaction between them; models that fall into this category are collectively referred to as conceptual models (Sayão 2001, 87).

It should be noticed that most of the definitions of “model” and “paradigm” raised within the scope of Information Science refer, directly or indirectly, to the concept proposed by Thomas Kuhn. The adoption of this perspective serves, for the most part, as an epistemological basis for the three theoretical and methodological perspectives in the field: physical, cognitive, and social, without, however, any evident concern for conceptually unravelling the terms “model” or “paradigm” themselves. This appropriation, often carried out functionally, favours a reading of paradigmatic transformations, while tending to neglect the methodological and representational specificities that distinguish models within each approach. In this context, Capurro (2010) makes a notable contribution by associating these three perspectives with a critique of technical-instrumental paradigms, emphasising the need for a hermeneutic and critical shift in Information Science that considers the cultural, ethical, and ontological mediation of information. Thus, it is understood that the use of the notion of paradigm in Information Science functions as theoretical support to justify changes in scientific orientation but rarely unfolds into more in-depth reflections on the theoretical-methodological status of the models used.

Table 2 presents the definitions of the authors studied on the notion of “model” and related terms.

¹² Reference of the original source cited: (Burt and Kinnucan 1990).

Table 2. Summary of the concept of “model”

Author	Disciplinary approach	Notion of “model”
Japiassú and Marcondes (1996)	Philosophy	“Paradigm, ideal form. Object that serves as a parameter for the construction or creation of others” (184).
Kuhn (1970; 2000)	Philosophy	The traditional idea of paradigms was that of models, “[...] particularly grammatical models of the right way to do things” (Baltas, Gavroglu, and Kindi 2000, 298), as well as the notion of exemplars within the structure and that of “hermeneutic basis for the science of a particular period” (Kuhn 2000, 221).
Hesse (1966)	Philosophy	“Let us agree that, when we speak of a model in its primary sense (in this discussion let us call it model ₁), we are not speaking of another object which can, as it were, be built or imagined alongside the phenomena we are investigating. The model ₁ is the imperfect copy (the billiard balls) minus the known negative analogy, so that we are only considering the known positive analogy, and the (probably open) class of properties about which it is not yet known whether they are positive or negative analogies” (9).
Abrantes (2020)	Philosophy	Scientific models are defined as fundamental tools for applying theories, explaining, predicting, and controlling phenomena. In addition, models play a crucial role in testing theories and simulating systems. “The term ‘model’ is ambiguous, being used by both scientists and philosophers with a plurality of meanings. An important methodological task is to categorise such models, show the contexts in which they are used, and for what purposes” (215).
Dutra (2021)	Philosophy	“This modelling activity is the basic, unified methodology of all scientific disciplines, in any field, as well as the methodology of everyday life. Models do in the domain of method what emergence does in the domain of ontology” (133).
Silva (2010)	Information Science	“[...] a model is a physical, mathematical or logical system that represents the essential structures of a reality and is capable, at its level, of dynamically explaining or reproducing the functioning of those same structures” (299).
Sayão (2001)	Information Science	He defines models as “simplified and intelligible representations of the world, which allow us to glimpse essential characteristics of a domain or field of study” (84), adding that they perform a “descriptive, explanatory, and, sometimes, observable function”, with a “heuristic dimension” that allows “the perception of other aspects not imagined before their elaboration” (83). He then presents the structure of information models as a continuum, based on the definition by Burt and Kinnucan (1990, cited in Sayão 2001, 87), categorising them as: cognitive models, focused on the user’s mind; conceptual models, linked to the system’s point of view; and semantic models, focused on representing the meaning of data.
Renault (2007)	Information Science	“Regarding models, in a more general sense, we can say that they can be employed in various aspects: epistemological, metaphysical, ethical, and aesthetic. However, we are interested in their epistemological meaning, which we understand as a way of explaining or representing reality. [...] First, we would like to make it clear that the concept of paradigm used here is established alongside theory, and its function is to serve as an umbrella under which various theories can be sheltered. [...]. As for models, although they presuppose theory, we could say that they are closer to method, to ‘how to do’ [...] It follows that the same paradigm can shelter several theories, just as a theory can shelter several models.” (54-5)

3. Discussion and results: possibilities and limitations of Hipátia as a model

Inspired by the philosopher Hypatia of Alexandria,¹³ the so-called Hipátia Model of Digital Preservation originated in 2018 as a result of a partnership between the Brazilian Institute of Information in Science and Technology (Ibict) and the Court of Justice of the Federal District and Territories (TJDFT), with a view to adopting a Trusted Digital Archival Repository (TDAR) to improve the Court's Records management process (Braga, Holanda, and Canelhas 2022; Pignataro et al. 2022; Braga 2022a; Brondani et al. 2024). Shintaku and Oliveira (2022, 69) explain it as follows:

Due to the importance of the philosopher Hypatia in science, Ibict, through its General Coordination of Information Technology (CGTI), named the archival digital preservation model after this scientist from the ancient world. Part of this model is BarraPres, a database capable of extracting, applying informational structures, and packaging data packets. Thus, BarraPres/Hipátia is free, open-source software developed and maintained by Ibict, designed to integrate systems following the Open Archival Information System (OAIS) model for preservation.

According to the literature review,¹⁴ studies addressing Hipátia are primarily focused on the digital preservation of records and information (Shintaku, Braga, and Oliveira 2021; Braga and Márdero Arellano 2022; Pignataro et al. 2023; Saad and Brito 2024). These works highlight, above all, its configuration as an TDAR (Braga et al. 2022; Pignataro, Menezes, and Braga 2022; Oliveira, Shintaku, and Silva 2022; Oliveira, Márdero Arellano, and Braga 2022; Braga 2022a; 2022b) and its operation (Braga 2022a; Shintaku and Oliveira 2022; Shintaku et al. 2023; Brondani et al. 2024).

In these works, the concept of Hipátia as a model for digital preservation of records and information predominates, except for Shintaku, Braga, and Oliveira (2021), who view it as software, a tool, or a system. The authors characterise it as a TDAR that complies with archival theory in relation to the chain of custody, a vision corroborated by other scholars (Pignataro, Menezes, and Braga 2022; Braga 2022a; Shintaku et al. 2023) and the authenticity of the preserved records (Braga, Holanda, and Canelhas 2022; Márdero Arellano 2022; Oliveira, Shintaku, and Silva 2022; Braga 2022a; 2022b; Pignataro et al. 2023; Brondani et al. 2024).

Regarding its operation, Braga (2022b, loc. 12) explains that “Hipátia operates based on a record-mass Producer.¹⁵ To date, no restrictions have been found regarding the scope of the Producer, which serves as an indication that Hipátia can adapt to different scenarios”. Its mediating role between the Open Archival Information System (OAIS) reference model and the different Producers for the preservation of their records and information can be observed in Shintaku, Braga, and Oliveira:

¹³ A personality who lived in Egypt during the Roman period. Her achievements in the fields of Mathematics, Physics, Astronomy, and Engineering have been widely documented and are still available (Shintaku, Braga, and Oliveira 2021).

¹⁴ The survey of bibliographic production on Hipátia considered the works available in the Hipátia Digital Collection (Ibict 2025), as well as works mapped in the Information Science Database (BRAPCI - Base de Dados em Ciência da Informação) on 10 March 2025. <https://brapci.inf.br/home>.

¹⁵ “Producer: The role played by those persons or client systems that provide the information to be preserved. This can include internal or external OAIS persons or systems” (CCSDS 2024, 15).

In this sense, the packaging process processes the metadata and files that are in the directory created by the crosswalk, forming a package in OAIS format. With this, the databus can deposit this package in Archivematica. Thus, **if other software meets the same OAIS model, its integration with the Hipátia archival databus can be carried out in a simple manner** (2021, 9; emphasis added).

Hipátia uses Archivematica as software for implementing TDAR, in accordance with the OAIS model. “Hipátia, developed and maintained by Ibict, is a system that functions as an automatic integrator of archival systems, receiving items from one system and packaging them for deposit in another, following the OAIS model” (Shintaku, Braga, and Oliveira 2021, 10).

The possibilities for improving and adapting Hipátia are outlined by Braga, Holanda, and Canelhas:

Hipátia was created to be an open model, constantly evolving. Its adoption by relevant institutions on the national scene suggests that **its conceptual basis has been accepted as a path for implementing TDAR**. Precisely because it is an open model, the incorporation of new elements becomes more fluid as various actors can contribute to this process. It is important to bear in mind the importance of establishing an evolution strategy for the Hipátia model, endorsing its avant-garde character while enabling a wide range of institutions to assert the pillars of preservation (Braga, Holanda, and Canelhas 2022, 10; emphasis added).

In this sense, its flexibility around interoperability (Shintaku et al. 2023) could, in terms of technological application, configure it as a model, at least in these terms:

Considering the diversity of systems proposed for managing records in digital format, it was established that the proposed model should enable its implementation in diverse information environments, regardless of the software and hardware used by the institution that intends to apply it. **This choice of a model with such comprehensive characteristics, although it requires greater effort in the model design process, also brings significant advantages by defining interoperability between technologies and information environments as a characteristic of the primary model** (Braga 2022a, 53-4; emphasis added).

The author then defends the adaptability of Hipátia as a model incorporated into a digital preservation project:

In addition to this definition of standards and technologies, another key requirement identified during the negotiating analysis of the aspects to be covered by the model concerns the possibility of customising the preservation process. Although there are regulations to be considered, as well as various guides and best practice guidelines, **since the intention is to allow the model to be applied to any information system that manages digital records**, there is also a need to enable the customisation of the preservation process to meet the specific needs of each institution. This customisation is important to ensure that fundamental aspects of the digital preservation project supported by the model are covered and that future access to the preserved content is guaranteed (Braga 2022a, 54; emphasis added).

Still on the specifics of applying Hipátia, we can cite Saad and Brito, who state that:

In this context, the Hipátia Model emerges as a system integration platform focused on digital preservation, **standing out for its flexibility and ability to integrate with various records Producers**. This model allows heterogeneous systems to communicate effectively, collecting and depositing data packages in destination systems, such as Archivematica, ensuring the reliable preservation of digital records (Saad and Brito 2024, 3; emphasis added).

The aforementioned characteristics of Hipátia seem to support its claim as a “model capable of adapting to any system that manages digital processes and objects” (Oliveira 2022, 8), as stated by Shintaku and Oliveira in relation to its databus, in a chapter of the same book: “[...] because it is prepared to follow strict archival standards, it is easy to use in other tools that follow the same standards” (Shintaku and Oliveira 2022, 75). Thus, the uniqueness and innovation of Hipátia seem to be linked to its databus: “The Hipátia databus acts as an integrative interface between systems, receiving, collecting, and depositing the packages in the destination system. This allows heterogeneous systems to establish communication, focused on digital preservation” (Oliveira, Shintaku, and Silva 2022, 88). These same authors explain that

Hipátia was designed to integrate with Archivematica, **but it allows for various Producers**, such as EIS [Electronic Information System] and the Electronic Judicial Process (PJe) used in the Court of Justice of the Federal District of Territories (TJDFT). As a result, it meets all system requirements, which, in turn, comply with international standards. **Thus, it can be utilized by various agencies or institutions, regardless of their archives processing system** (Oliveira, Shintaku, and Silva 2022, 82; emphasis added).

Although most of the studies analysed understand it as a model, the characteristics that justify this have not been clearly presented. Thus, we decided to verify, according to the theoretical references of Philosophy and Information Science contained in the previous section, the possibilities and limitations of characterising Hipátia as a model for the digital preservation of records and information.

The definition of the concept of model, by Japiassú and Marcondes (1996), as a paradigm, can be linked to Kuhn’s (1970) concept of paradigm. In the sense of parameter, according to the former authors, it is believed that Hipátia could not be seen as a model. However, in the sense of exemplars, according to Kuhn (1970), Hipátia could provide guidelines for solving problems related to the digital preservation of records and information.

From Hesse’s perspective (1966), Hipátia could also be characterised as a model, an analogy of OAIS in the phenomenon of digital preservation, which provides for its operation through BarraPres and other modules that support the proposal. Thus, it could function as a scientific model, through analogies between known and unknown systems, interfering with the behaviour of complex systems based on their structural similarities with other systems and digital preservation models. It is important to highlight the flexibility of the model in the face of demands surrounding the understanding of the phenomenon, a characteristic that is evident in the functioning of Hipátia.

Amid the ambiguity recognised by Abrantes (2020) and among the types of models listed by the scholar, Hipátia can be viewed both as an analogue model, given its relationship with the OAIS

reference model and the possibility of its application with different Producers, and as a representation of a real system, which, in this case, concerns its interoperability with Producers.

In the relationship that Dutra (2021) establishes between the model and the method, Hipátia could also take a model approach, in which it has OAIS as a reference model for the digital preservation of records and information, serving as a method in the application of that model for the implementation of an TDAR.

Likewise, taking Renault's study (2007) as a reference, from an epistemological perspective, Hipátia could be understood as a model in terms of applicability, based on a theory represented by the OAIS reference model. In this sense, when operationalising the implementation of TDAR, in accordance with the OAIS recommendations, Hipátia would be "[...] then the reality – actual or supposed – that the theory [OAIS reference model] seeks to explain" (Ferrater Mora 2004, 1989, cited in Renault 2007, 54).¹⁶

Moreover, it is worth revisiting Sayão's (2001) concept, according to which models perform descriptive, explanatory, and heuristic functions, i.e., they not only describe a phenomenon, but also make it intelligible and open to evolution and adaptation. The author states that models can be classified, within a continuous configuration, according to their proximity to the user, the system, or the interaction between the two. Thus, Hipátia can be understood as a model and, complementarily, as a "data model," since it approximates the definition of information systems and their technical and functional structures by operating in the integration between Producers and preservation systems, complying with regulatory standards and promoting interoperability between different technological environments.

According to Silva's study (2010), Hipátia could be a logical model representing the record and information preservation structures of the past, that is, those previously dedicated to the preservation of traditional records and which are now reproduced for the preservation of digital records and information (born-digital and digitised records, in addition to their metadata), taking into account the similarities, differences and singularities of the two structures. In addition to the information management model mentioned by the scholar, there are also management models (within the scope of Producers) and preservation models, in which Hipátia operates. Following the example of the model presented by him – eLit.pt – Hipátia diagnoses preservation demands, represents them, and intervenes, as an applied method of OAIS in interface with Archivematica, via BarraPres.

4. Final considerations

The OAIS model postulates essential requirements for the managerial variables that must be considered by organizations undertaking digital preservation initiatives. It does so by identifying the agents involved and defining their respective roles and responsibilities, as well as by classifying the types of information exchanged among these agents. Nevertheless, the model has a referential nature, and for organizations it remains a technological and operational challenge to implement these requirements in order to ensure long-term preservation (Rosenthal et al. 2005).

¹⁶ Reference to the original cited: (Ferrater Mora 2004)

The Hipátia Model, as suggested by the theoretical model analyzed in this study, originated with the aim of operationalizing the functional requirements established by the OAIS model, thereby supporting the implementation of digital preservation strategies by different institutions. Would flexibility, adaptability, interoperability, and reproducibility be sufficient to characterise Hipátia as a model for the digital preservation of records and information? Would the same be true based on its innovative features in relation to its databus?

Taking into account the theoretical references of Philosophy and Information Science studied, Hipátia would not be configured as a model according to the definition of a Philosophy dictionary, but could be characterised from the perspective of an example (aspect of the notion of paradigm linked to model); an analogy of another model (in this case, OAIS, applying it in the context of digital preservation); a method in the model view, again serving as an application of OAIS; and a logical model, which represents other preservation structures (previously physical and now logical).

From the perspective of OAIS as a reference model, from which Hipátia appropriates guidelines for the digital preservation of records and information, it can be seen as an application of that model, that is, a method for implementing a TDAR. This methodological conception does not limit it; on the contrary, it links it to the reference model and enhances the possibilities of its uses in different contexts, through its characteristics of flexibility, adaptability, interoperability, and reproducibility.

Based on the literature review undertaken, it was possible to observe that the notion of model presents itself as a complex, multifaceted epistemological construct, permeated by different theoretical traditions. From the perspective of Philosophy and Information Science, it was found that models fulfil explanatory, descriptive, and operational functions, serving both to represent and intervene in the phenomena they seek to understand.

When applying these benchmarks to the analysis of Hipátia, it was identified that the system has characteristics that qualify it as a model in multiple senses. It can be understood as a model in the sense that Kuhn attributes to exemplary, as an analogical model in the light of Hesse, as an operational representation according to Abrantes, and as a methodological instrument linked to a theory, according to Dutra and Renault. Its function of integration and preservation between heterogeneous systems reinforces its power as an applied model, especially when considering its adherence to OAIS principles and its capacity for diagnosis and intervention, as also analysed by Silva.

In view of the above, it can be concluded that Hipátia can indeed be considered a model for the digital preservation of records and information, provided that a broader and multidimensional understanding of the notion of “model” is adopted. Its relationship with OAIS, its methodological structure based on integration between systems, and its diagnostic and operational function show that Hipátia goes beyond being mere software, taking on a representative and intervening role in digital archival processing. By integrating theory, method, and practice, it contributes not only to the technical consolidation of TDAR, but also to the epistemological maturation of digital preservation in the field of Archival Science and Information Science. Its potential lies precisely in the articulation between its normative bases and its adaptability to the complexity of institutional contexts, which effectively makes it a model with theoretical relevance and practical applicability. Furthermore, it is important to emphasize that digital preservation should not be understood

merely as the implementation of technological solutions, but as a continuous management process involving organizational, technical, and security dimensions. Studies on the long-term preservation of electronic records highlight that the reliability and authenticity of records depend on the establishment of procedural rules embedded in recordkeeping and record-preservation systems and integrated with institutional business processes, ensuring systematic control over the creation, handling, and preservation of records (Duranti 2001; Vermaaten, Lavoie, and Caplan 2012). In addition, preservation environments must ensure the maintenance of the identity and integrity of digital records over time, which requires knowledge of the record's elements and the documentation of any changes occurring during preservation processes (Duranti 2001). At the same time, digital preservation strategies must consider the risks that threaten the long-term accessibility and usability of digital materials, since successful preservation depends on identifying and mitigating threats that may compromise essential preservation properties (Vermaaten, Lavoie, and Caplan 2012). In this context, preservation initiatives must articulate organizational policies, technological infrastructures, and risk-management strategies capable of sustaining trustworthy digital preservation environments and ensuring the continued accessibility, authenticity, and integrity of digital records over time (Rosenthal et al. 2005).

References*

- Abrantes, Paulo C. 2020. *Método e ciência: uma abordagem filosófica*. Belo Horizonte: Fino Traço.
- Baltas, Aristides, Kostas Gavroglu, and Vassiliki Kindi. 2000. "A discussion with Thomas S. Kuhn." In *The road since structure: Philosophical Essays, 1970-1993, with an Autobiographical Interview*, edited by Thomas S. Kuhn, 256-323. Chicago: University of Chicago Press.
- Braga, Tiago Emmanuel Nunes. 2022a. "O modelo Hipátia: a proposta do Ibict para a preservação digital arquivística." In *Hipátia: modelo de preservação para repositórios arquivísticos digitais confiáveis*, edited by Tiago Emmanuel Nunes Braga and Miguel Á. Márdero Arellano, 52-65. Brasília: Ibict. <https://bit.ly/braga-modelo-hipatia>.
- Braga, Tiago Emmanuel Nunes. 2022b. "Hipátia: a primeira experiência prática de implementação de um RDC-Arq no Brasil." Paper presented at the Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação, Porto Alegre, November 7-11. <https://bit.ly/braga-hipatia-primeira-experiencia>.
- Braga, Tiago Emmanuel Nunes, Alex Pereira de Holanda, and Tatiana Canelhas. 2022. "Resolução RDC-Arq Conarq: uma análise dos novos requisitos informacionais propostos." *Revista Brasileira de Preservação Digital* 3: 1-10. <https://bit.ly/braga-holanda-canelhas>.
- Braga, Tiago Emmanuel Nunes, Miguel Ángel Márdero Arellano. 2022. *Hipátia: modelo de preservação para repositórios arquivísticos digitais confiáveis*. Brasília: Ibict. <https://bit.ly/braga-hipatia-rd-carq>.
- Brondani, Marcelo Moro, Denise Oliveira de Araújo, Milton Shintaku, and Alexandre Faria de Oliveira. 2024. "Validação de documentos arquivísticos digitalizados: aplicação de tecnologia para apoiar a garantia de confiabilidade de arquivos preservados." Paper presented at the Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação, Vitória, ES. November 4-8. <https://enancib.ancib.org/index.php/enancib/xxivenancib/paper/view-File/2861/1529>.
- Burt, Patricia V., and Kinnucan, Mark T. 1990. "Information models and modeling techniques for information systems." In *Annual review of Information Science and Technology*, edited by Martha E. Williams, 175-208. Silver Spring, MD: American Society for Information Science.
- Capurro, Rafael. 2010. "Epistemología y ciencia de la información." *Acimed* 21 (2): 248-65. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352010000200008.
- CNPQ (Conselho Nacional de Desenvolvimento Científico e Tecnológico). n.d. *Tabela de Áreas do Conhecimento*. <https://lattes.cnpq.br/documents/11871/24930/TabeladeAreasdoConhecimento.pdf>.
- CCSDS (Consultative Committee for Space Data Systems). 2024. *Reference Model For An Open Archival Information System (OAIS)*. Washington, DC: CCSDS Secretariat. <https://ccsds.org/Pubs/650x0m3.pdf>.

* Last date of access to websites: January 23, 2026.

Cunha, Murilo Bastos da, and Cordélia Robalinho de Oliveira Cavalcanti. 2008. *Dicionário de Biblioteconomia e Arquivologia*. Brasília: Briquet de Lemos. <http://repositorio.unb.br/handle/10482/34113>.

Domingues, Ivan. 2004. *Epistemologia das ciências humanas. Tomo I: Positivismo e Hermenêutica. Durkheim e Weber*. São Paulo: Edições Loyola.

Duranti, Luciana 2001. "The long-term preservation of authentic electronic records".

Proceedings of the 27th VLDB Conference. https://www.interpares.org/display_file/ip1-2_dissemination_cpr_duranti_vldb_2001.pdf.

Dutra, Luiz Henrique. 2021. "O que a realidade uniu, que o filósofo não separe! Modelos nas ciências da natureza e da sociedade." In *Perspectivas sobre o método científico*, edited by Jorge Alberto Molina, 115-36. Seropédica, RJ: PPGFIL-UFRRJ.

Eugênio, Marconi, Ricardo Orlandi França, and Rui Campos Perez. 1996. "Ciência da Informação sob a ótica paradigmática de Thomas Kuhn: elementos de reflexão." *Perspectivas em Ciência da Informação* 1 (1): 27-39.

González Teruel, Aurora. 2005. *Los estudios de necesidades y usos de la información: fundamentos y perspectivas actuales*. Gijón: Trea.

Hesse, Mary B. 1966. *Models and analogies in science*. Notre Dame, Indiana: University of Notre Dame Press.

Ibict (Instituto Brasileiro de Informação em Ciência e Tecnologia). 2025. *Hipátia: acervo digital*. <https://bit.ly/ibict-hipatia-acervo-digital>.

Japiassú, Hilton, and Danilo Marcondes. 1996. *Dicionário básico de Filosofia*. Rio de Janeiro: Jorge Zahar.

Japiassú, Hilton, and Danilo Marcondes. 2001. *Dicionário básico de Filosofia*. Rio de Janeiro: Jorge Zahar.

Kuhn, Thomas S. 2000. *The road since structure: Philosophical Essays, 1970-1993, with an Autobiographical interview*. Chicago: University of Chicago Press.

Kuhn, Thomas S. 1970. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Nehmy, Rosa Maria Quadros, Carlos Henrique Rezende Falci, Jarbas Gregue Acosta, and Rosane R. Fraga. 1996. "A ciência da informação como disciplina científica." *Perspectivas em Ciência da Informação* 1 (1): 9-25.

Oliva, Alberto. 1994. "Kuhn: o normal e o revolucionário na reprodução da racionalidade científica." In *Filosofia, História e Sociologia das Ciências I: abordagens contemporâneas*, edited by Vera Portocarrero, 67-102. Rio de Janeiro: FIOCRUZ.

Oliveira, Alexandre Faria, Miguel Ángel Márdero Arellano, and Tiago Emmanuel Nunes Braga. 2022. "A visão do Ibict para um ecossistema nacional de preservação digital." In *Hipátia: modelo de preservação para repositórios arquivísticos digitais confiáveis*, edited by Tiago Em-

manuel Nunes Braga and Miguel Ángel Márdero Arellano, 91-100. Brasília: Ibict. <https://doi.org/10.22477/9786589167501.cap7>.

Oliveira, Alexandre Faria, Milton Shintaku, and M. S. Silva. 2022. “Casos de sucesso de implementação do Modelo RDCArq e Hipátia.” In *Hipátia: modelo de preservação para repositórios arquivísticos digitais confiáveis*, edited by Tiago Emmanuel Nunes Braga and Miguel Ángel Márdero Arellano, 79-90. Brasília: Ibict. <https://doi.org/10.22477/9786589167501.cap6>.

Oliveira, Cecília Leite. 2022. “Apresentação.” In *Hipátia: modelo de preservação para repositórios arquivísticos digitais confiáveis*, edited by Tiago Emmanuel Nunes Braga and Miguel Ángel Márdero Arellano. Brasília: Ibict. <https://doi.org/10.22477/9786589167501>.

Oliveira, Roberto Cardoso. 1997. *Tempo e tradição: interpretando a Antropologia. Sobre o pensamento antropológico*. Rio de Janeiro: Tempo Brasileiro.

Otlet, Paul. 1934. *Traité de Documentation: le livre sur le livre, theorie et pratique*. Bruxelles: Mundaneum.

Pignataro, Tatiana Canelhas, Cristiano Menezes, and Tiago Emmanuel Nunes Braga. 2022. “Os principais normativos relacionados ao RDCArq.” In *Hipátia: modelo de preservação para repositórios arquivísticos digitais confiáveis*, edited by Tiago Emmanuel Nunes Braga and Miguel Ángel Márdero Arellano, 42-53. Brasília: Ibict. <https://doi.org/10.22477/9786589167501.cap3>.

Pignataro, Tatiana Canelhas, Marília Cristina da Silva, Alex Pereira de Holanda, Sérgio Augusto Coelho Diniz Nogueira Júnior, Marcos Sigismundo da Silva, Ívina Flores Melo, and Marilete da Silva Pereira. 2023. “A gestão dos Formatos de preservação de Arquivos no Archivematica: Estudo de Caso com o Processo Judicial Eletrônico.” *Anais do Workshop de Informação, Dados e Tecnologia - WIDaT 6*. <https://doi.org/10.22477/vi.widat.51>.

Renault, Leonardo Vasconcelos. 2007. “Paradigmas e modelos: proposta de análise epistemológica para a Ciência da Informação.” *Informação e Sociedade* 17 (2): 53-60. <https://periodicos.ufpb.br/index.php/ies/article/view/636>.

Renault, Leonardo Vasconcelos. 2023. *Currículo Lattes*. Brasília, DF: Conselho Nacional de Desenvolvimento Científico e Tecnológico. <http://lattes.cnpq.br/2611007188909877>.

Rosenthal, David S. H, Thomas Robertson, Tom Lipkis, Vicky Reich, and Seth Morabito 2005. “Requirements for digital preservation systems: a bottom-up approach”. *D-Lib Magazine* 11(11). <https://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html>.

Saad, Rondineli Gama, and Eva Juani Santos Brito. 2024. “Implementação do modelo Hipátia no Tribunal de Justiça do Distrito Federal e Territórios (TJDFT): um estudo sobre maturidade e práticas de preservação digital.” *Revista Brasileira de Preservação Digital* 5. <https://brapci.inf.br/v/314509>.

Schulz, Peter. 2023. “Thomas Kuhn, uma breve crônica de sua trajetória e de seus paradigmas.” *Problemata - Revista Internacional de Filosofia* 14 (4): 7-14.

Sayão, Luis Fernando. 2001. “Modelos teóricos em ciência da informação: abstração e método científico.” *Ciência da Informação* 30 (1). <https://brapci.inf.br/v/18262>.

Sayão, Luis Fernando. 2025. *Currículo Lattes*. Brasília, DF: Conselho Nacional de Desenvolvimento Científico e Tecnológico. <http://lattes.cnpq.br/3422623122948389>.

Shintaku, Milton, Tiago Emmanuel Nunes Braga, and Alexandre Faria Oliveira. 2021. “Hipátia: uma ferramenta livre no apoio à preservação digital.” *Revista Brasileira de Preservação Digital* 2: E021006. [10.20396/rebpred.v2i00.15987](https://doi.org/10.20396/rebpred.v2i00.15987).

Shintaku, Milton, and Alexandre Faria Oliveira. 2022. “Arquitetura do Barrapres/Hipátia.” In *Hipátia: modelo de preservação para repositórios arquivísticos digitais confiáveis*, edited by Tiago Emmanuel Nunes Braga and Miguel Ángel Márdero Arellano, 66-78. Brasília: Ibict. <https://doi.org/10.22477/9786589167501.cap5>.

Shintaku, Milton, Alexandre Faria Oliveira, Lucas Ângelo Silveira, Diego José Macêdo, and Ingrid Torres Schiessl. 2023. “Uso do protocolo de interoperabilidade OAI-ORE para preservação digital.” *Revista Brasileira de Preservação Digital*, 4: 1-16. <https://brapci.inf.br/v/225162>.

Silva, Armando Malheiro da. 2010. “Modelos e Modelizações em Ciência da Informação: o Modelo eLit.pt e a investigação em literacia informacional.” *Prisma.Com* 13. <https://ojs.letras.up.pt/index.php/prismacom/article/view/2011>.

Silva, Sérgio Conde de Albite, Julio César Cardoso, and Alex da Silva Gomes de Brito. 2005. “Entrevista com Armando Malheiro da Silva.” *Arquivística.net* 1 (1). <https://brapci.inf.br/index.php/res/v/50957>.

Silva, Zayr Claudio Gomes da, and Jonathas Luiz Carvalho Silva. “O conceito de “paradigma” de Thomas Khun: interesses e controvérsias na Ciência da Informação.” Paper presented at the Encontro Nacional de Pesquisa em Ciência da Informação, Aracaju, November 6-10, 2023. <https://ancib.org/enancib/index.php/enancib/xxxiiienancib/paper/view/1452/1004>.

Vermaaten, Sally, Brian Lavoie, and Caplan Priscilla. 2012. “Identifying threats to successful digital preservation: the SPOT model for risk assessment”. *D-Lib Magazine* 18(9/10). <https://www.dlib.org/dlib/september12/vermaaten/09vermaaten.html>.

Publication times and costs in Open Access journals: A thematic analysis of the Scopus–DOAJ ecosystem*

Jean Paul Castillo Nuñez^(a)
Lloy Pinedo-Tuanama^(b)
Paolo Arones-Perez^(c)
Madeleine Gavilán-Colca^(d)

a) Universidad Privada Norbert Wiener, <https://orcid.org/0000-0001-6052-5472>

b) Universidad Privada Norbert Wiener, <https://orcid.org/0000-0002-5569-8739>

c) Universidad Privada Norbert Wiener, <https://orcid.org/0000-0002-7086-6438>

d) Universidad Privada Norbert Wiener, <https://orcid.org/0009-0007-1475-1015>

Contact: Jean Paul Castillo Nuñez, jean.castillo@uwiener.edu.pe; Lloy Pinedo-Tuanama, lloy.pinedo@uwiener.edu.pe;
Paolo Arones-Perez, paolo.arones@uwiener.edu.pe; Madeleine Gavilán-Colca, a2023200322@uwiener.edu.pe

Received: 14 December 2025; **Accepted:** 19 March 2026; **First Published:** 15 May 2026

ABSTRACT

The expansion of open access has positioned Article Processing Charges (APCs) and publication times as key dimensions for understanding the functioning of the scholarly publishing system. This study aims to characterise both components in scientific journals that are simultaneously indexed in Scopus and listed in DOAJ, identifying patterns by subject area and quartile. Data from Scopus, the 2024 SJR and DOAJ were integrated, considering only active journals with available information on average time to publication and APCs. The analysis, descriptive in nature, examined the distribution of these indicators across subject areas and editorial levels. The results show that publication times are similar among journals within the same subject area but differ when comparing different subject areas, whereas publication costs display larger differences in the upper quartiles and tend to decrease in the lower ones, with parallel trajectories in Life Sciences, Physical Sciences and Health Sciences and a differentiated pattern in Social Sciences. The study provides an updated empirical characterisation of how temporal stability and economic stratification coexist within a broad segment of the open access ecosystem.

KEYWORDS

Open access; APCs; Publication timelines; Editorial processes; Subject area disparities.

* Jean Paul Castillo-Nuñez es autor de: Conceptualización, Curación de datos, Análisis formal, Investigación, Metodología, Supervisión, Visualización, Redacción – borrador original, Redacción – revisión y edición; Lloy Pinedo-Tuanama es autor de: Curación de datos, Análisis formal, Software, Metodología, Validación, Redacción – revisión y edición; Paolo Arones-Pérez es autor de: Conceptualización, Curación de datos, Redacción – borrador original; Madeleine Gavilán-Colca es autora de: Conceptualización, Curación de datos, Redacción – borrador original.

© 2026, The Author(s). This is an open access article, free of all copyright, that anyone can freely read, download, copy, distribute, print, search, or link to the full texts or use them for any other lawful purpose. This article is made available under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. JLIS.it is a journal of the SAGAS Department, University of Florence, Italy, published by EUM, Edizioni Università di Macerata, Italy, and FUP, Firenze University Press, Italy.

Tiempos y costos de publicación en revistas de acceso abierto: una lectura temática del ecosistema Scopus–DOAJ

RESUMEN

La expansión del acceso abierto ha situado los Cargos por Procesamiento de Artículos (APC) y los tiempos de publicación como dimensiones clave para comprender el funcionamiento del sistema editorial. Este estudio tiene como objetivo caracterizar ambos componentes en revistas científicas simultáneamente indexadas en Scopus y registradas en DOAJ, identificando patrones por área temática y cuartil. Se integraron datos de Scopus, SJR 2024 y DOAJ, considerando únicamente revistas activas con información disponible sobre tiempo promedio hasta la publicación y APC. El análisis, de carácter descriptivo, examinó la distribución de estos indicadores por disciplinas y niveles editoriales. Los resultados muestran que los tiempos de publicación son similares entre revistas de una misma área temática, pero difieren al comparar áreas temáticas distintas, mientras que los costos de publicación presentan diferencias más amplias en los cuartiles superiores y tienden a reducirse en los inferiores, con trayectorias paralelas en Life Sciences, Physical Sciences y Health Sciences y un comportamiento diferenciado en Social Sciences. El estudio ofrece una caracterización empírica actualizada de cómo coexisten estabilidad temporal y estratificación económica en un segmento amplio del ecosistema de acceso abierto.

PALABRAS CLAVE

Acceso abierto; APC; Tiempos de publicación; Proceso editorial; Desigualdad temática.

1. Introducción

Seleccionar una revista científica adecuada para publicar resultados de investigación constituye una de las decisiones más críticas para académicos e instituciones, especialmente en el actual ecosistema de la comunicación científica. Esta elección no solo determina la visibilidad y el alcance de los hallazgos, sino que también influye en la trayectoria profesional de los investigadores y en la reputación de las universidades (Björk y Solomon 2015). Tradicionalmente, los criterios predominantes han girado en torno a indicadores de impacto como el Journal Citation Reports o el Scimago Journal Rank (SJR), ampliamente incorporados en políticas institucionales y sistemas de evaluación científica (Mingers y Leydesdorff 2015). En este contexto, la selección de revistas se ha convertido también en una práctica estratégica para los investigadores, en la que intervienen consideraciones de visibilidad, reconocimiento académico y adecuación temática, más allá del simple cumplimiento de criterios formales de calidad (Reyes Rodríguez y Moraga Muñoz 2020). No obstante, estos indicadores, aunque útiles, no capturan otras dimensiones fundamentales de la actividad editorial ni los costos asociados al modelo de publicación.

En este marco, es importante definir que el acceso abierto constituye distintos modelos, entre los cuales están el green open access, basado en el autoarchivo sin cargos para los autores; el gold open access, financiado habitualmente mediante cargos por procesamiento de artículos (APC); el modelo híbrido, que permite el pago de APC en revistas por suscripción; y el diamond o platinum open access, sin costos para autores ni lectores (Tennant y Lomax 2019; Costello 2019; Meagher 2021; White et al. 2021; Fallahhusein et al. 2025). Esta heterogeneidad implica que las dinámicas económicas del acceso abierto no son uniformes, sino dependientes del esquema editorial adoptado.

En particular, dentro de este ecosistema del acceso abierto, el incremento sostenido de la producción científica en las últimas décadas ha generado nuevas tensiones en el sistema editorial, entre ellas el aumento de los APC y la presión por acelerar los tiempos de aceptación y publicación. En el contexto del acceso abierto, estas tensiones se intensifican, Pallares et al. (2022) y Soto-Herrera, Beigel, y Pallares (2025), muestran que, si bien el modelo abierto contribuye a democratizar el acceso al conocimiento, también traslada gastos significativos hacia autores e instituciones, lo que configura barreras económicas especialmente marcadas en países del Sur Global. En particular, Soto-Herrera, Beigel, y Pallares (2025) reportan que investigadores de Argentina y Colombia afrontaron APC promedio de 1 593 y 2 112 dólares por artículo entre 2013 y 2020. Además, la distribución del acceso abierto sigue siendo desigual, como evidencia el hecho de que solo el 1% de los artículos registrados en el Directory of Open Access Journals (DOAJ) provienen de países de bajos ingresos (Escobar Espinoza 2023). Desde una perspectiva estructural, los APC han sido identificados como un mecanismo de estratificación dentro del acceso abierto, al introducir barreras económicas diferenciadas según áreas temáticas y contextos institucionales (Klebel y Ross-Hellauer 2023). A ello se suman desafíos vinculados con la ética editorial, la diversidad en la autoría y la rigurosidad de la revisión por pares, cuestiones que han sido objeto de creciente atención en la literatura reciente (Raudales-García et al. 2024).

El análisis del modelo económico del acceso abierto muestra, además, una elevada heterogeneidad en los APC que no puede explicarse únicamente por los costos editoriales, sino que se asocia a factores como el área temática, el tipo de editor y el posicionamiento editorial de las

revistas registradas en DOAJ (Siler y Frenken 2020). La concentración del mercado editorial, el tamaño de los editores y la adopción de modelos híbridos de acceso abierto son factores clave en la determinación de los precios de los APC (Morrison et al. 2017; Budzinski et al. 2020). Asimismo, el nivel de impacto de las revistas se ha identificado como un componente relevante en la configuración de los APC, con diferencias claras entre revistas de acceso abierto pleno y modelos híbridos (Schönfelder 2020). Esta dinámica también se manifiesta en variaciones entre áreas temáticas. Diversos estudios han señalado que los cargos por procesamiento de artículos (APC) no se distribuyen de manera uniforme entre disciplinas, sino que tienden a ser más elevados en varios campos científicos en comparación con las ciencias sociales (Gray 2020). Desde la perspectiva de los autores, estas diferencias se materializan en cargas económicas muy desiguales, particularmente visibles en áreas como la medicina, donde los costos acumulados por publicación pueden alcanzar montos elevados incluso a nivel individual (Ellingson et al. 2021). Si bien dentro del ecosistema DOAJ coexisten revistas que operan sin cargos por publicación, el presente estudio se centra específicamente en el segmento de revistas que aplican APC, con el fin de analizar las dinámicas económicas asociadas a los modelos de publicación basados en cargos por procesamiento de artículos.

En este escenario, las revistas simultáneamente indexadas en Scopus y registradas en DOAJ constituyen un subconjunto particularmente relevante para observar cómo se configuran las dinámicas editoriales y económicas del acceso abierto. Perdomo, More, y Flores (2023) y Perdomo y Morales (2022), analizan, en áreas como comunicación y arquitectura, la diversidad de políticas editoriales y la existencia de una correlación entre los costos de APC y el cuartil de las revistas. Solomon y Björk (2012), por su parte, muestran que, al controlar variables como disciplina y antigüedad, las tasas de citación pueden ser similares entre revistas de acceso abierto y de suscripción, aunque las revistas OA que aplican APC tienden a recibir un mayor promedio de citas. En el contexto latinoamericano, persisten problemas como información editorial desactualizada y retrasos en los procesos de revisión, factores que pueden afectar la reputación de las revistas y su visibilidad internacional, incluso en el modelo OA (González-Sanabria et al. 2020; Sánchez-Tarragó et al. 2016). En conjunto, estos trabajos abordan dimensiones parciales de la relación entre políticas editoriales, APC, prácticas de comunicación y posicionamiento, pero no ofrecen aún una visión integrada de los patrones estructurales que se observan en el cruce Scopus–DOAJ.

Los tiempos editoriales constituyen otra dimensión clave de estas dinámicas. Los plazos de evaluación y publicación varían significativamente según el área temática y los tiempos reales suelen ser más prolongados que los reportados por las revistas en DOAJ (Zabala et al. 2023). Empero, dado que esta información proviene de autodeclaraciones editoriales, su interpretación requiere cautela metodológica, especialmente cuando se emplea con fines comparativos. En algunos casos, como el de la pandemia de COVID-19, se ha observado una aceleración temporal de los procesos editoriales, aunque este fenómeno no representa un cambio estructural generalizable en el sistema (Horbach 2020).

En este marco, el presente estudio describe patrones editoriales en un conjunto de revistas científicas de OA indexadas en Scopus y registradas en DOAJ, examinando la distribución de sus tiempos de publicación y APC, segmentados por cuartiles y áreas temáticas según la clasificación SJR. El objetivo es ofrecer una caracterización estructural de este subconjunto de revistas, identificando tendencias que permitan contextualizar mejor las dinámicas del OA en términos de

accesibilidad económica, eficiencia editorial y posicionamiento dentro del sistema de indización. La pregunta de investigación que orienta este análisis es: ¿qué patrones se observan en la distribución de los APC y los tiempos editoriales dentro de las revistas de acceso abierto incluidas en Scopus y DOAJ, cuando se analizan por cuartil y área temática? Este enfoque aporta una perspectiva descriptiva sobre aspectos centrales del ecosistema de publicación en acceso abierto y ofrece elementos útiles para investigadores, editores y responsables de políticas académicas interesados en comprender las dinámicas estructurales que configuran la sostenibilidad y la equidad del modelo OA.

2. Metodología

El estudio se desarrolló bajo un enfoque cuantitativo y transversal, orientado a describir patrones editoriales en revistas científicas de acceso abierto incluidas simultáneamente en Scopus y en el DOAJ. El objetivo metodológico fue caracterizar la distribución de los tiempos editoriales y los APC en este subconjunto de revistas, segmentando los resultados según el posicionamiento por cuartiles y las áreas temáticas establecidas por Scopus.

La base de datos se construyó a partir del archivo oficial Scopus Title List Content de Elsevier exportado el 06 de junio de 2025, permitiendo identificar todas las revistas con estatus activo, a la fecha de corte. A este registro se integraron los indicadores de SJR 2024, utilizados exclusivamente para clasificar las revistas por cuartiles y por área temática conforme al esquema All Science Journal Classification (ASJC). Adicionalmente, se añadió a partir del título de la revista la información declarada en DOAJ a la misma fecha de corte, que incluye los tiempos promedio de publicación (en semanas) y los valores de APC expresados en dólares estadounidenses; estos valores fueron estandarizados utilizando el tipo de cambio promedio del primer semestre de 2025.

El proceso de selección consideró únicamente revistas que cumplieran cuatro condiciones simultáneas: (1) estar activas en Scopus; (2) figurar en DOAJ; (3) contar con un valor de SJR disponible y (4) con APC (costo declarado mayor a 0). La restricción a valores positivos respondió a la necesidad de analizar la variabilidad económica del modelo basado en cargos por publicación a partir de una distribución monetaria continua. La inclusión de revistas con APC igual a cero habría introducido una concentración significativa de valores nulos, afectando la interpretación de las medidas de tendencia central y dispersión y dificultando la comparación entre niveles de posicionamiento editorial.

Se excluyeron las revistas que, aun estando indexadas, no disponían de SJR, lo que puede ocurrir por su reciente incorporación, un número insuficiente de citas o porque no alcanzan los umbrales mínimos requeridos para el cálculo del indicador.

De esta forma, para cada revista, se organizó la siguiente información: área temática SJR, cuartil, tiempo promedio hasta publicación y valor del APC. El tratamiento estadístico, mediante SPSS en su versión 26, incluyó el cálculo de medidas de tendencia central (media y mediana) y de dispersión (desviación estándar y coeficiente de variación) tanto a nivel global como segmentado por cuartil. El coeficiente de variación se empleó como indicador estandarizado de heterogeneidad relativa, permitiendo comparar la dispersión entre variables de distinta magnitud y entre segmentos con diferentes promedios. Adicionalmente, se analizaron posibles asimetrías mediante el

contraste entre media y mediana, complementado con la inspección de valores extremos y rangos intercuartílicos.

Los resultados se presentaron segmentados por cuartil y por las cuatro grandes áreas del esquema ASJC (Life Sciences, Social Sciences, Physical Sciences y Health Sciences), con el propósito de identificar patrones estructurales y diferencias interdisciplinarias dentro del ecosistema Scopus–DOAJ. A partir de esta estructura se efectuó un análisis descriptivo, mostrando las distribuciones segmentadas por cuartil y área temática en las cuatro secciones principales del ASJC (Life Sciences, Social Sciences, Physical Sciences y Health Sciences). Este enfoque permitió identificar patrones y diferencias entre campos, sin buscar establecer relaciones causales entre las variables. En cuanto a limitaciones, es importante señalar que los tiempos editoriales (desde la recepción hasta la publicación) y los APC provienen de la información declarada por cada revista en DOAJ, lo que puede implicar variaciones en los criterios de reporte entre editores. Asimismo, el análisis se restringió exclusivamente a revistas con valor de SJR disponible y con APC, lo que excluye títulos potencialmente relevantes que aún no cumplen los requisitos de citación necesarios para su cálculo o que operan bajo el modelo de acceso abierto diamante, respectivamente.

3. Resultados

La población total de revistas registradas en Scopus a junio de 2025 asciende a 47 637 títulos, de los cuales 31 068 permanecen activas y 16 569 están inactivos. Es decir, aproximadamente dos tercios del conjunto se mantienen en operación, mientras que un tercio corresponde a títulos que han dejado de publicar o han sido discontinuados. Dentro del segmento activo, 6355 revistas figuran en el DOAJ, lo que representa cerca de una quinta parte de las revistas activas en Scopus (Figura 1). Este subconjunto delimita el espacio específico en el que convergen, de manera simultánea, indización en una base de datos de amplio alcance y adhesión explícita al modelo de OA.

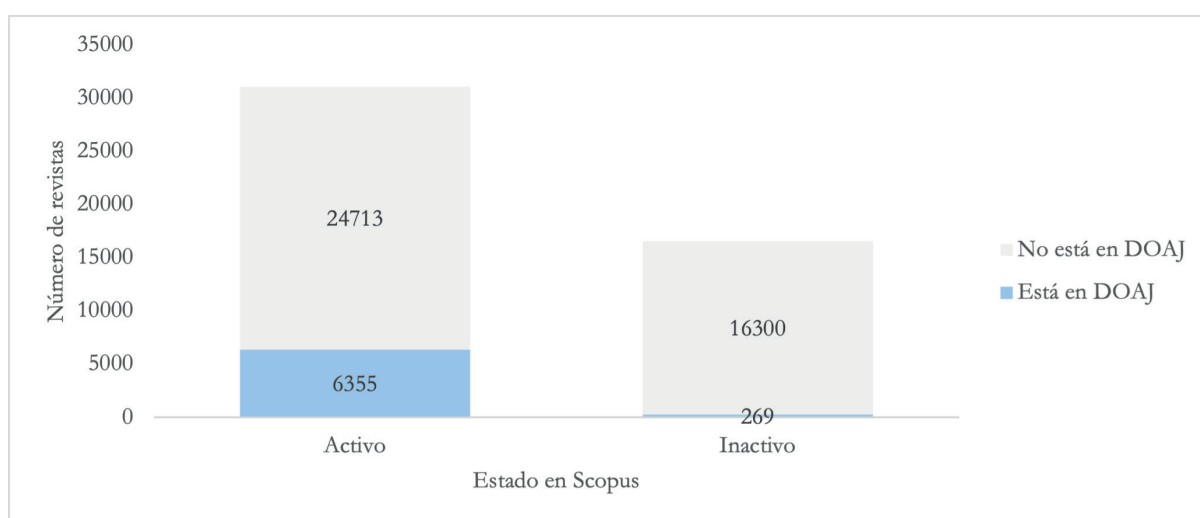


Figura 1. Revistas según estado en Scopus y registro en DOAJ

De las revistas que cumplen ambas condiciones (Scopus–DOAJ), 212 no disponen de un valor de SJR, ya sea por su reciente incorporación al índice o por un volumen de citación todavía insuficiente para el cálculo del indicador, quedando 6143 revistas. Dentro de este conjunto, 3 584 revistas (58%) declaran aplicar cargos por procesamiento de artículos ($APC > 0$), mientras que 2 559 (42%) no reportan costos de publicación (Figura 2). Esta distribución muestra que, aún dentro del cruce Scopus–DOAJ con métrica disponible, el modelo basado en APC constituye el segmento mayoritario, aunque coexiste con una proporción relevante de revistas que operan sin cargos directos a los autores.

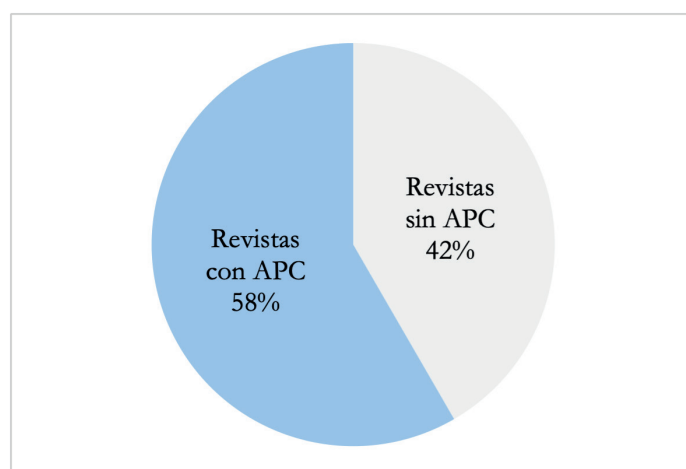


Figura 2. Proporción de revistas de acceso abierto con y sin APC

El análisis de costos y tiempos editoriales se realizó sobre las 3 584 revistas que declaran APC mayores a cero, cuya base detallada se puede visualizar en el Suplementario 1. Esta delimitación permite examinar la variabilidad de los montos dentro del modelo basado en cargos por publicación, manteniendo la coherencia con el diseño analítico adoptado.

Con el fin de caracterizar la estructura general del conjunto antes de su desagregación por cuartiles y áreas temáticas, se calcularon estadísticos descriptivos globales para ambas variables (Tabla 1). Este análisis inicial permite dimensionar el comportamiento agregado de los tiempos editoriales y los APC dentro del ecosistema Scopus–DOAJ con APC declarado.

La tabla 1 evidencia que el tiempo promedio hasta la publicación fue de 15,81 semanas, con una mediana de 14 semanas y un coeficiente de variación de 58,25%, indicando una dispersión moderada en los procesos editoriales dentro del conjunto analizado. La desviación estándar de 9,21 semanas confirma la existencia de diferencias entre revistas, aunque concentradas en un rango relativamente estable en torno al promedio.

En contraste, los APC presentaron una media de 1 829,90 USD y una mediana de 1 800 USD, con una desviación estándar de 1 152,87 USD y un coeficiente de variación de 63%, evidenciando una mayor heterogeneidad económica entre las revistas. El coeficiente de variación superior al observado en los tiempos editoriales sugiere que la dimensión económica del modelo de publicación exhibe una variabilidad relativa más acentuada que la dimensión temporal del proceso editorial.

Tabla 1. Distribución de los tiempos de publicación y APC en revistas Scopus–DOAJ

Variable	Media	Mediana	Desviación estándar	Coefficiente de variación (%)
Tiempo hasta publicación (semanas)	15,81	14,00	9,21	58,25
APC (USD)	1829,90	1800,00	1152,87	63,00

En términos comparativos, mientras los tiempos editoriales tienden a agruparse en torno a un intervalo relativamente acotado, los APC presentan una dispersión proporcionalmente mayor, lo que anticipa la existencia de diferencias estructurales más marcadas en la dimensión económica que en la operativa. A partir de este panorama agregado, se procede a examinar la distribución de ambas variables según cuartil y área temática, con el fin de identificar patrones diferenciados dentro del sistema.

3.1. Tiempos de publicación

Los tiempos editoriales constituyen uno de los indicadores operativos más relevantes del proceso de comunicación científica. Su duración influye en la oportunidad con la que la investigación se incorpora al debate académico, pero también refleja prácticas, flujos y dinámicas internas de cada área del conocimiento.

Con el fin de explorar la relación entre posicionamiento bibliométrico e intensidad temporal del proceso editorial, se examinó la asociación entre el valor continuo del índice SJR y el tiempo promedio hasta publicación. El diagrama de dispersión (Figura 3) muestra una alta concentración de revistas en los rangos bajos y medios del SJR, con una dispersión vertical relativamente amplia en ese tramo y una reducción progresiva de la variabilidad en los valores más altos del índice. El coeficiente de determinación obtenido fue extremadamente bajo ($R^2 = 0,002$), lo que indica que el impacto editorial explica una proporción prácticamente nula de la variación observada en los tiempos de publicación. En términos descriptivos, este resultado sugiere que la duración del proceso editorial se comporta de manera prácticamente independiente del posicionamiento bibliométrico de la revista. Aunque la línea de tendencia presenta una pendiente levemente negativa, su magnitud es mínima y no configura un patrón estructural consistente. Asimismo, la distribución de los puntos evidencia que revistas con alto y bajo impacto pueden compartir tiempos editoriales similares, lo que refuerza la idea de que la dimensión temporal no opera como un mecanismo claro de estratificación dentro del ecosistema analizado.

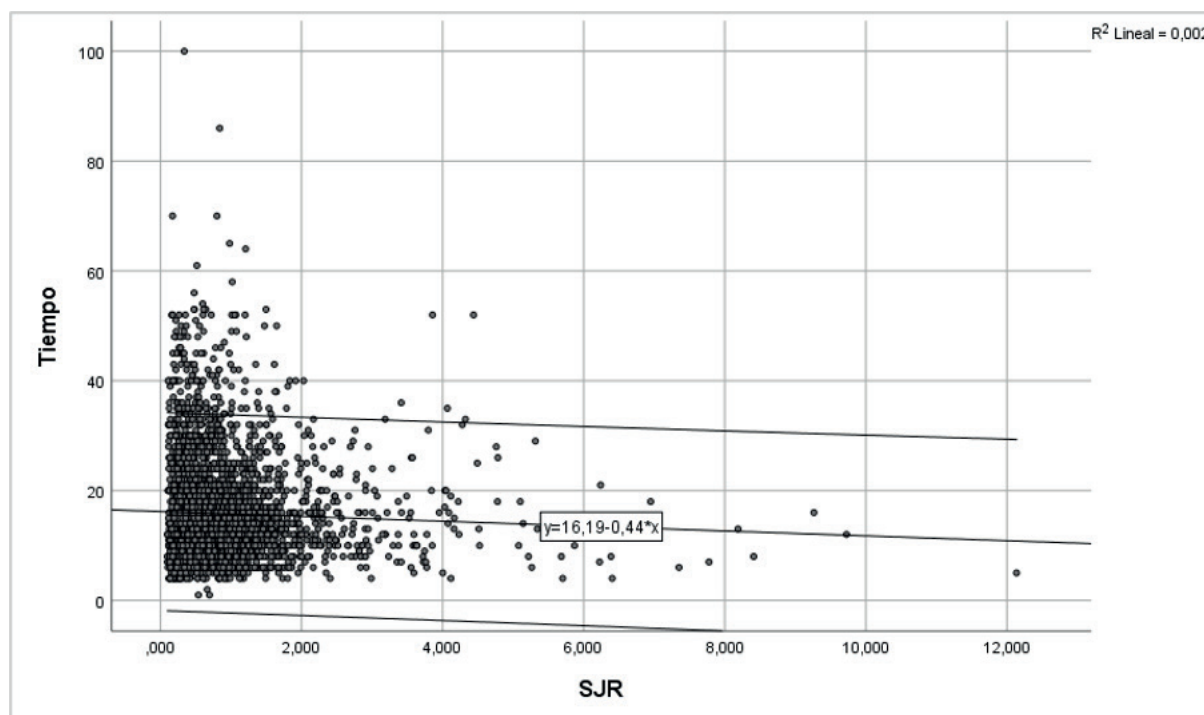


Figura 3. Diagrama de dispersión entre SJR y tiempo promedio hasta publicación en revistas Scopus–DOAJ.

La distribución segmentada por cuartil confirma esta estabilidad, como se muestra en la Tabla 2. Las medias oscilan entre 15,51 semanas en Q2 y 16,65 semanas en Q3, mientras que las medianas permanecen constantes en 14 semanas para Q1, Q2 y Q3, con una ligera reducción a 13 semanas en Q4. La estabilidad de la mediana, en contraste con variaciones menores en la media, sugiere que las diferencias observadas entre cuartiles están influenciadas principalmente por valores extremos y no por modificaciones sustantivas en la duración típica del proceso editorial. De igual manera, los coeficientes de variación se mantienen en un rango estrecho, entre 56,13% y 61,98%, lo que indica una dispersión relativa moderada y homogénea entre niveles de posicionamiento. La consistencia de estos valores refuerza la interpretación de que la variabilidad interna del tiempo editorial es comparable en revistas de distinto cuartil, sin evidencia de segmentación temporal progresiva conforme se desciende en el ranking.

Tabla 2. Distribución de los tiempos de publicación por cuartil SJR en revistas Scopus–DOAJ.

Estadístico	Cuartil			
	Q1	Q2	Q3	Q4
Media	15,66	15,51	16,65	15,98
Mediana	14,00	14,00	14,00	13,00
Desviación estándar	8,79	9,00	10,32	9,53
Coeficiente de variación (%)	56,13	58,03	61,98	59,64

El análisis por áreas temáticas introduce una diferenciación más visible entre disciplinas que entre cuartiles. A partir de la Figura 4 se examinan los promedios de semanas transcurridas desde la recepción del manuscrito hasta su publicación final en revistas de acceso abierto indexadas en Scopus y registradas en DOAJ, distinguiendo los patrones observables en cuatro grandes áreas temáticas.

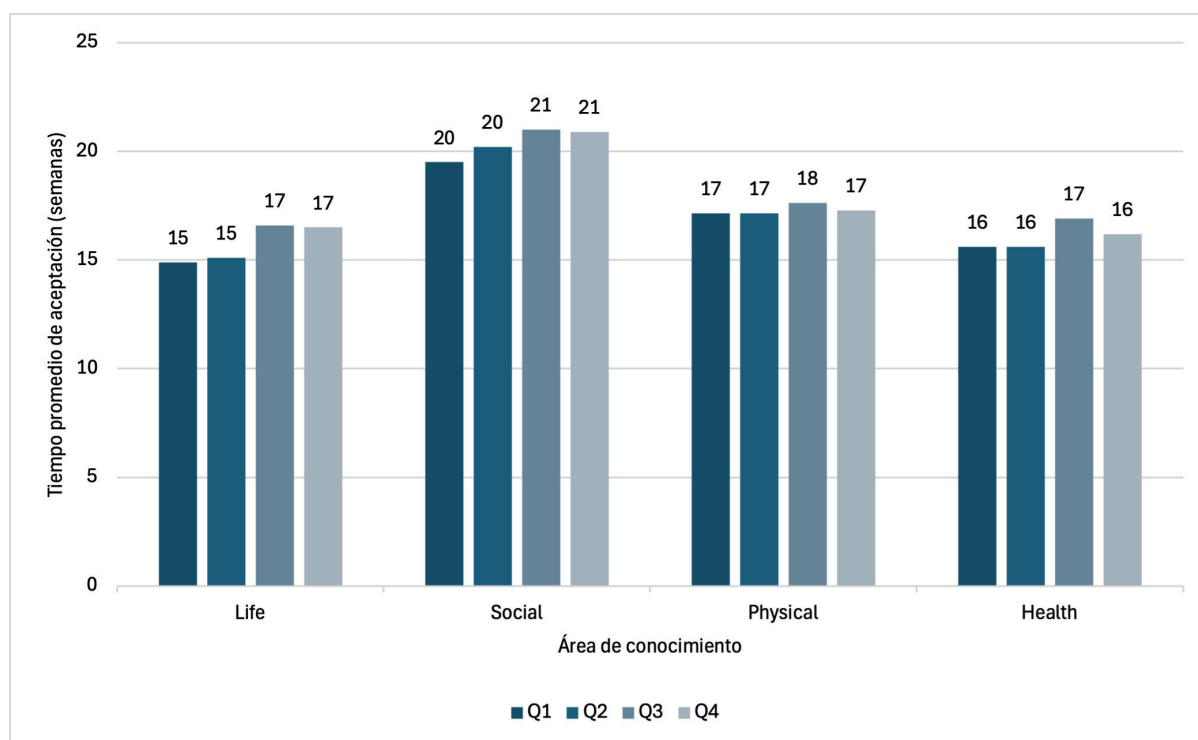


Figura 4. Promedio de tiempo hasta publicación (en semanas) por cuartil y área temática de revistas de acceso abierto en Scopus y DOAJ.

En Life Sciences, los tiempos se ubican consistentemente en torno a las 15 semanas para Q1 y Q2, con un aumento a 17 semanas en Q3 y Q4. Aunque la diferencia entre cuartiles es moderada, destaca la presencia de un ciclo editorial relativamente breve y estable, lo que sugiere procesos bien estructurados en términos de revisión y producción.

En Social Sciences, el comportamiento es distinto: los plazos más prolongados del conjunto se concentran aquí, con un mínimo de 20 semanas y un máximo de 21 semanas. Esta concentración en valores más altos puede asociarse a prácticas disciplinares que habitualmente requieren evaluaciones más extensas y mayor diversidad de revisores, sin que ello implique diferencias marcadas entre cuartiles dentro del área.

En Physical Sciences, los tiempos se sitúan en un margen estrecho entre 17 y 18 semanas. Esta regularidad sugiere un flujo editorial más homogéneo, posiblemente vinculado a rutinas de procesamiento técnico relativamente estables y a una menor variabilidad interna en los tipos de manuscritos enviados a estas revistas.

El caso de Health Sciences presenta una de las distribuciones más simétricas del análisis: 16 semanas en Q1, Q2 y Q4, y un leve incremento a 17 semanas en Q3. Esta uniformidad refuerza la idea

de que algunas áreas científicas mantienen ritmos editoriales claramente estandarizados y poco sensibles a diferencias de posicionamiento entre revistas.

En consecuencia, las diferencias más marcadas se observan entre áreas (interdisciplinarias) más que entre cuartiles (intradisciplinarias), lo que indica que la duración del proceso editorial parece estar más asociada a dinámicas propias de cada campo que al nivel de impacto de la revista dentro de ese campo.

Finalmente, el examen por grupo editorial, presentado en la Figura 5, muestra variaciones organizacionales más amplias que las detectadas entre cuartiles. Wiley y Taylor & Francis registran los tiempos promedio más prolongados (18 semanas), seguidos por Springer Nature y el grupo “Otros” (17 semanas). Elsevier presenta un promedio intermedio (16 semanas), mientras que SAGE (14 semanas) y Frontiers Media SA (13 semanas) se sitúan por debajo del conjunto. MDPI exhibe el tiempo promedio más reducido (7 semanas), configurando el valor más bajo del grupo analizado. La amplitud observada entre el mínimo y el máximo editorial supera claramente la variación detectada entre cuartiles, lo que sugiere que las diferencias organizacionales pueden tener un peso mayor en la duración declarada del proceso editorial que el posicionamiento bibliométrico en sí mismo.

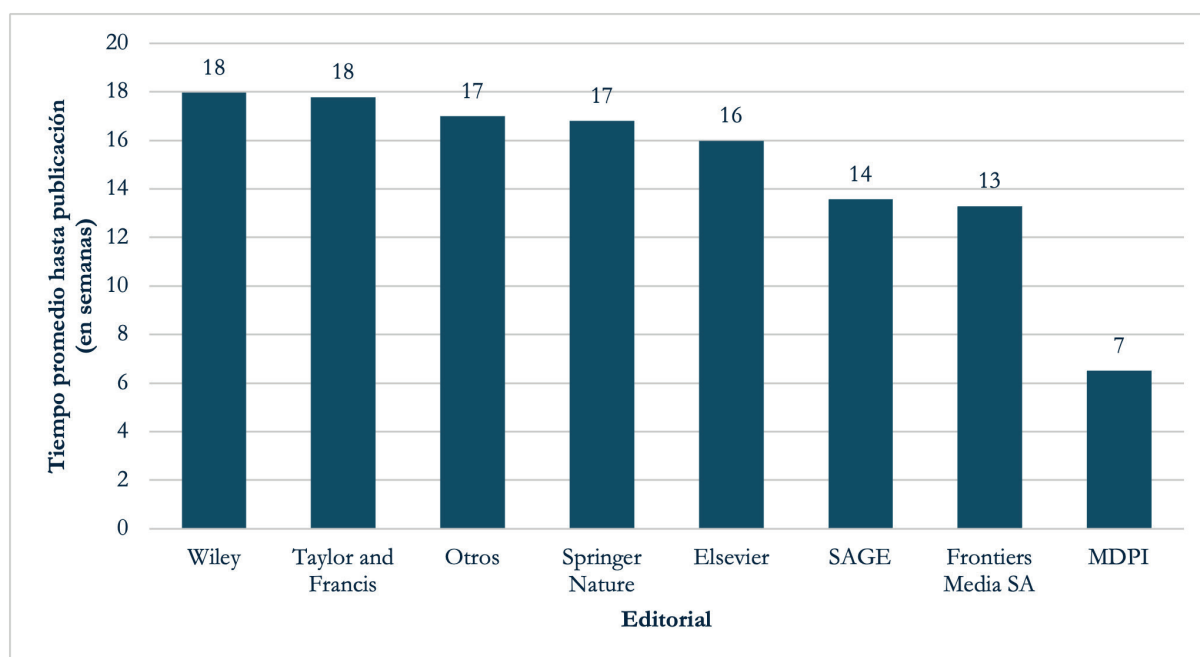


Figura 5. Tiempo promedio hasta publicación (en semanas) según grupo editorial.

En conjunto, los resultados muestran que la dimensión temporal del acceso abierto indexado en Scopus y registrado en DOAJ presenta una estructura relativamente estable en términos de cuartiles, con diferencias acotadas y niveles de dispersión comparables, mientras que las variaciones más perceptibles emergen al contrastar áreas temáticas y grupos editoriales. Esta configuración descriptiva permite caracterizar el tiempo editorial como una variable con baja dependencia del impacto y mayor sensibilidad a factores disciplinares y organizacionales.

3.2. Costos de publicación

Los APC constituyen un componente central del modelo de publicación en acceso abierto, pues representan la principal vía de financiamiento para un número creciente de revistas. A diferencia de la dimensión temporal, cuya variabilidad resultó limitada frente al impacto bibliométrico, la estructura de costos exhibe patrones más marcados de estratificación económica.

Con el propósito de examinar la relación entre posicionamiento editorial y magnitud de los cargos declarados, se analizó la asociación entre el valor continuo del SJR y el monto del APC. El diagrama de dispersión correspondiente se presenta en la Figura 6. A diferencia de lo observado en los tiempos de publicación, la nube de puntos evidencia una tendencia positiva respaldada por un coeficiente de determinación bajo pero significativo en términos descriptivos ($R^2 = 0,160$). Este resultado indica que una proporción relevante de la variación en los APC se asocia con el nivel de impacto de la revista.

La Figura 6 muestra una marcada concentración de revistas en niveles de APC relativamente elevados dentro de la escala representada, con una aglomeración particularmente visible a partir del umbral de los 1 000 USD. Esta concentración refleja que una proporción importante de revistas del conjunto analizado aplica cargos de publicación en rangos altos dentro del modelo del acceso abierto con APC.

Al mismo tiempo, la amplia dispersión vertical de los puntos indica que revistas con niveles similares de impacto pueden presentar diferencias sustantivas en los cargos declarados, lo que sugiere la coexistencia de múltiples estrategias tarifarias dentro de segmentos comparables por factor de impacto.

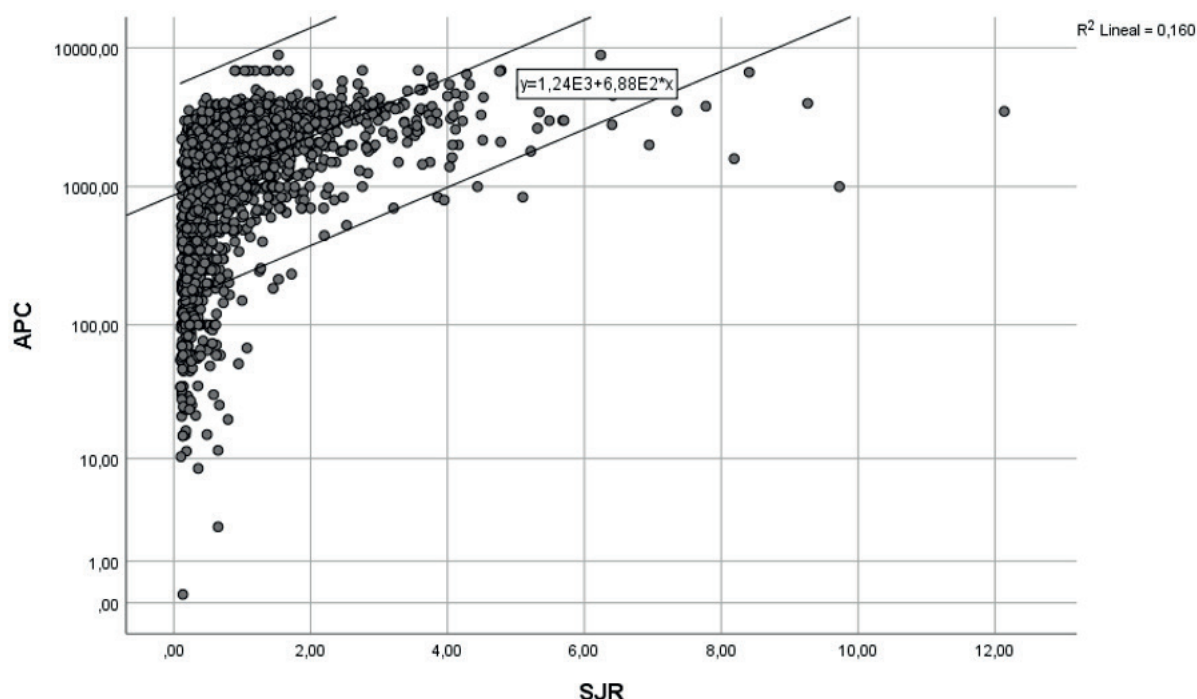


Figura 6. Diagrama de dispersión entre SJR y APC en revistas Scopus-DOAJ.

La segmentación por cuartil refuerza este patrón estructural. La Tabla 3 muestra una reducción progresiva de los montos promedio conforme se descende en el ranking: la media pasa de 2 349,08 USD en Q1 a 543,21 USD en Q4. Esta trayectoria decreciente no solo refleja una diferenciación clara entre niveles de impacto, sino que también revela cambios en la dispersión relativa. El coeficiente de variación aumenta de 47,58% en Q1 a 111,92% en Q4, lo que indica que las revistas de menor posicionamiento presentan mayor heterogeneidad en sus modelos de cobro. Asimismo, la brecha entre media y mediana se amplía en los cuartiles inferiores —especialmente en Q4—, sugiriendo distribuciones más asimétricas y la coexistencia de estrategias económicas diversas dentro de ese segmento.

Tabla 3. Distribución de los APC por cuartil SJR

Estadístico	Cuartil			
	Q1	Q2	Q3	Q4
Media	2349,08	1772,98	1106,01	543,21
Mediana	2305,00	1790,00	900,00	298,66
Desviación estándar	1117,72	990,76	870,87	607,96
Coeficiente de variación (%)	47,58	55,88	78,74	111,92

El análisis por áreas temáticas introduce una dimensión adicional a esta estratificación. La Figura 7 sintetiza los valores promedio de APC en revistas indexadas en Scopus y registradas en DOAJ, distinguidos por área temática y segmentación en cuartiles. Esta organización permite observar cómo se distribuyen los costos dentro del ecosistema OA, identificando tendencias estructurales relevantes.

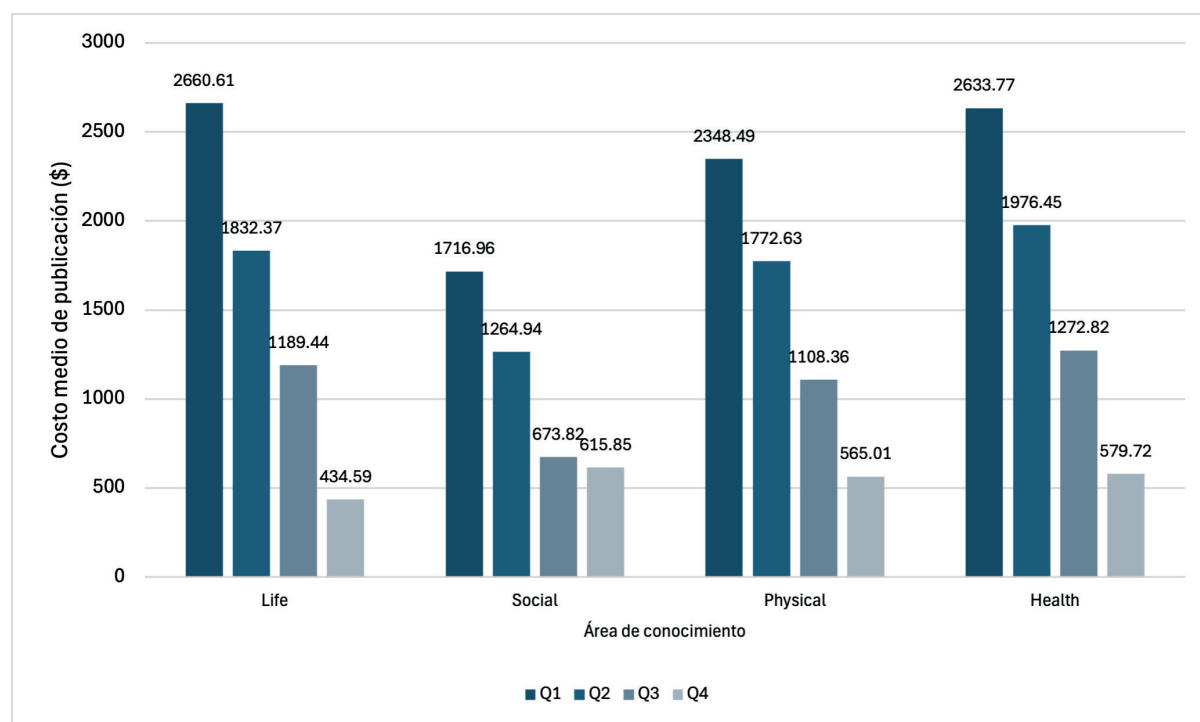


Figura 7. Promedio de APC (\$) por cuartil y área temática de revistas de acceso abierto en Scopus y DOAJ.

En el cuartil superior del sistema (Q1), los APC alcanzan sus valores más altos y configuran un primer contraste entre campos. Life Sciences, Health Sciences y Physical Sciences presentan montos cercanos entre sí, 2660.61, 2633.77 y 2348.49 USD, consolidando un bloque de áreas caracterizado por costos elevados. Social Sciences, en comparación, registra un valor considerablemente menor (1716.96 USD), lo que la posiciona como el área relativamente más accesible dentro del segmento mejor situado del sistema. Así, aun compartiendo el mismo nivel de posicionamiento, las revistas Q1 no forman un conjunto homogéneo: la carga económica para los autores varía sustancialmente entre áreas.

Al observar el siguiente nivel (Q2), los APC descienden de manera perceptible y las distancias entre áreas comienzan a estrecharse. La jerarquía general, con Life, Health y Physical Sciences por encima de Social Sciences, permanece, pero ya sin las brechas amplias que caracterizan al cuartil superior. Q2 actúa así como un segmento intermedio donde las diferencias entre disciplinas siguen presentes, aunque con una dispersión menos marcada, y donde el sistema económico del acceso abierto muestra una primera tendencia hacia la moderación.

En el tramo central del sistema editorial (Q3), la reducción de costos es aún más pronunciada. Los valores se agrupan en un rango más compacto y la separación entre áreas se atenúa de forma clara. Este cuartil funciona como un punto de transición: las revistas continúan exhibiendo diferencias internas entre campos, pero la magnitud de estas diferencias se vuelve más contenida, lo que revela un segmento en el que las condiciones económicas del acceso abierto tienden a armonizarse. El cuartil inferior (Q4) muestra el punto de mayor convergencia en los APC. Life Sciences, Physical Sciences y Health Sciences registran montos próximos, entre 434 y 580 USD, evidenciando una notable proximidad en sus modelos de cobro en el extremo más accesible del sistema. Social Sciences, sin embargo, rompe parcialmente esta convergencia: con 615.85 USD, se convierte en el área más costosa en Q4, invirtiendo la posición relativa que ocupaba en Q1. Aunque la diferencia es pequeña en términos absolutos, esta inversión revela un patrón singular dentro del área, cuya estructura económica se comporta de forma distinta a la del resto de áreas en el cuartil inferior.

En conjunto, los resultados muestran que la distribución de APC responde a una estructura escalonada: los dos primeros cuartiles concentran los costos más altos y exhiben las diferencias más amplias entre áreas, mientras que los cuartiles inferiores se caracterizan por montos reducidos y distancias menores entre disciplinas. Dentro de este panorama, Life Sciences, Physical Sciences y Health Sciences comparten trayectorias paralelas, con gradientes amplios entre sus segmentos superiores e inferiores; en contraste, Social Sciences sigue una ruta diferenciada, con un Q1 comparativamente más accesible y un Q4 relativamente más elevado. Esta combinación destaca que las variaciones económicas del acceso abierto no solo dependen del nivel de posicionamiento editorial, sino también de las configuraciones propias de cada campo temático.

Por último, el examen por grupo editorial permite identificar diferencias organizacionales en la fijación de precios. La Figura 8 muestra que Frontiers Media SA registra el promedio más elevado (2 901,91 USD), seguida por Springer Nature (2 653,16 USD). Un segundo conjunto lo integran Taylor & Francis (2 246,07 USD), Wiley (2 239,99 USD) y SAGE (2 191,88 USD), con valores próximos entre sí. MDPI (2 078,41 USD) y Elsevier (1 949,82 USD) se sitúan por debajo de este grupo, mientras que la categoría agregada de “Otros” presenta un promedio significativamente menor (861,19 USD).

La amplitud entre el máximo y el mínimo editorial supera ampliamente la variación observada entre algunos cuartiles intermedios, lo que evidencia que la estructura de precios también se encuentra asociada a dinámicas organizacionales y modelos de negocio diferenciados. En términos descriptivos, los grandes conglomerados comerciales tienden a concentrarse en los rangos superiores de APC, mientras que editoriales de menor escala o menor presencia relativa operan en segmentos de menor costo.

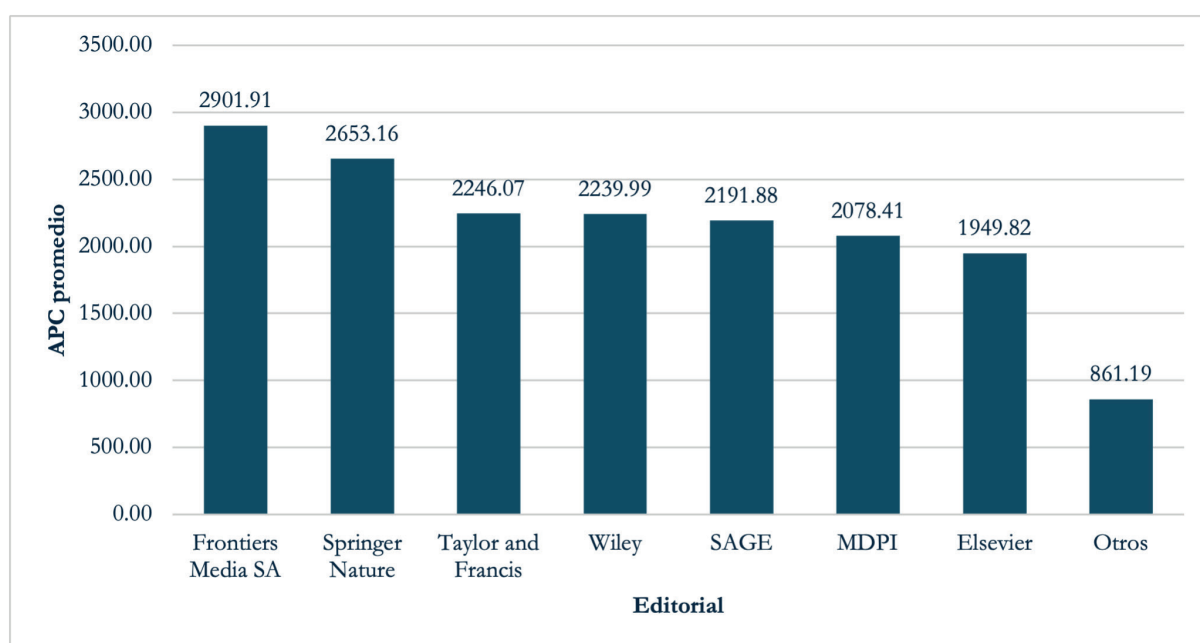


Figura 8. APC promedio según editorial

En conjunto, los resultados muestran que, a diferencia de la dimensión temporal, la estructura de los APC presenta una relación más consistente con el posicionamiento bibliométrico y una diferenciación económica más marcada tanto entre cuartiles como entre áreas y conglomerados editoriales. Esta configuración permite caracterizar los costos de publicación como una variable con gradiente jerárquico visible y sensibilidad tanto disciplinar como organizacional dentro del ecosistema Scopus–DOAJ.

4. Discusiones y conclusión

El cruce Scopus–DOAJ permite observar, con una escala poco frecuente, cómo se articulan dos dimensiones operativas del acceso abierto: la temporalidad del proceso editorial y la estructura económica asociada a los APC. En este universo, el modelo basado en cargos por publicación no solo concentra una fracción mayoritaria de las revistas con SJR disponible, sino que además exhibe patrones internos diferenciados: mientras la dimensión temporal se mantiene relativamente estable frente al posicionamiento bibliométrico, la dimensión económica muestra una estratificación más marcada por cuartil, disciplina y conglomerado editorial. En este sentido, los resultados del

presente estudio permiten examinar de manera integrada cómo ambas dimensiones se configuran dentro del universo de revistas indexadas simultáneamente en Scopus y registradas en DOAJ.

En términos temporales, el conjunto analizado ($APC > 0$) se caracteriza por tiempos promedio hasta publicación de 15,81 semanas y una mediana de 14 semanas, con una dispersión moderada ($CV = 58,25\%$). Esta centralidad es consistente con la idea de que la duración declarada del ciclo editorial, en agregados amplios, tiende a organizarse en rangos operativos relativamente acotados. Los resultados obtenidos en este estudio muestran que, al contrastar cuantitativamente con estudios disciplinares, se aprecia que varios campos reportan intervalos más extensos: en neurocirugía se observó un promedio de 207 días desde envío hasta publicación ($\approx 29,6$ semanas) (Jubran et al. 2023), y en atención primaria un promedio de 243,4 días ($\approx 34,8$ semanas) (Chen et al. 2024). En genética, los valores medianos de envío a publicación se situaron entre 137 y 166 días ($\approx 19,6$ – $23,7$ semanas), dependiendo del cuartil (Zimmer et al. 2023). Estas diferencias, aun considerando que provienen de muestras y diseños distintos, delimitan un marco comparativo útil: el segmento Scopus–DOAJ con APC muestra tiempos promedio más bajos y una estructura más homogénea, sin que ello implique por sí mismo una lectura normativa sobre calidad.

Dentro del propio conjunto, el patrón más consistente es la baja dependencia del tiempo editorial respecto del impacto. La asociación entre SJR continuo y tiempo promedio es prácticamente nula ($R^2 = 0,002$), y por cuartiles las medias varían estrechamente ($\approx 15,5$ – $16,7$ semanas), con medianas estables (≈ 13 – 14 semanas). Estos resultados sugieren que, dentro del universo analizado, la duración del proceso editorial no se organiza de forma lineal en función del posicionamiento bibliométrico de las revistas. Este comportamiento coincide con la evidencia de que, aun cuando existen diferencias relevantes en “turnaround” entre revistas, estas no necesariamente se ordenan de manera lineal por métricas de impacto. En ciencias pesqueras, por ejemplo, se reportó que las medianas de tiempo a aceptación oscilaron entre 64 y 269 días ($\approx 9,1$ – $38,4$ semanas) y las medianas de tiempo a publicación entre 79 y 323 días ($\approx 11,3$ – $46,1$ semanas), con distribuciones típicamente sesgadas a la derecha y amplitudes internas considerables por revista (Runde 2021). En conjunto, estos antecedentes refuerzan la interpretación de que, en el caso analizado, la temporalidad responde con mayor peso a dinámicas disciplinares y organizacionales que a una jerarquía bibliométrica en sí misma, en línea con las cautelas sobre la interpretación directa de métricas como indicadores de calidad editorial (Wilches-Visbal y Castillo-Pedraza 2022; Tort, Targino, y Amaral 2012).

A diferencia de la temporalidad, los APC exhiben una estructura económica claramente escalonada y más sensible al posicionamiento. El promedio global (1 829,90 USD) y la mediana (1 800 USD) se acompañan de una heterogeneidad elevada ($CV = 63\%$), que se amplifica al descender en cuartiles: las medias pasan de 2 349,08 USD en Q1 a 543,21 USD en Q4, mientras la variabilidad relativa se incrementa hasta 111,92% en el cuartil inferior. Los resultados del presente estudio indican que este patrón es consistente con la relación parcial entre impacto y precio reportada en estudios previos sobre APC, donde el costo no se explica únicamente por costos técnicos sino por posicionamiento editorial y segmentación del mercado (Björk y Solomon 2015; Siler y Frenken 2020; Budzinski et al. 2020). En el plano microestructural, la relación positiva entre SJR y APC en el conjunto analizado ($R^2 = 0,244$) indica una asociación moderada: el impacto contribuye a ordenar los precios, pero coexisten múltiples estrategias tarifarias dentro de rangos similares de SJR, lo que mantiene una dispersión sustantiva.

La comparación cuantitativa con evidencia externa refuerza esa lectura. En revistas quirúrgicas exclusivamente open access, se observó un costo mediano de 100 USD (rango 0–2 580), con 47% de títulos sin cargos y 13% por encima de 2 000 USD; además, las correlaciones lineales entre costo y métricas fueron débiles (Yuen, Muquit, y Whitfield 2018). En ingeniería y tecnología, los APC oscilaron entre 7,6 y 3 471,5 USD, con un promedio de 727 USD, y se reportó una correlación positiva débil entre APC e impacto (Ezema 2019). En un análisis centrado en MDPI, se encontró una correlación más fuerte entre APC e indicadores de estatura ($R^2 \approx 0,619$ con CiteScore y $R^2 \approx 0,64$ con Impact Factor) y un rango de cargos entre 0 y 2 000 CHF (Okagbue, Teixeira da Silva, y Anake 2020). Estas comparaciones permiten contextualizar los resultados del presente estudio, mostrando que la fuerza de la asociación impacto-precio depende del subconjunto analizado (multieditorial vs. un solo publisher) y del campo, lo que es coherente con el hecho de que el modelo APC no opera como un esquema uniforme, sino como un arreglo heterogéneo de precios y políticas.

Por áreas temáticas, los resultados reproducen diferencias consistentes con patrones descritos en DOAJ. En Q1, Life Sciences, Health Sciences y Physical Sciences concentran APC más elevados ($\approx 2\,348$ – $2\,661$ USD), mientras Social Sciences se ubica por debajo ($\approx 1\,717$ USD), lo que confirma configuraciones económicas disciplinares diferenciadas dentro del mismo nivel de posicionamiento. En un análisis de publishers con DOAJ Seal, se reportó que más del 28% de títulos cobraban más de 1 500 USD y cerca del 5% más de 2 000 USD, además de identificar medicina como el campo con mayor concentración de títulos caros (Rodrigues, Abadal, y Araújo 2020). En consonancia con estos antecedentes, los resultados del presente estudio muestran que los campos biomédicos y de ciencias de la vida tienden a sostener niveles superiores de APC en los segmentos de mayor impacto, mientras que la trayectoria de Social Sciences presenta un comportamiento diferenciado al invertirse parcialmente en Q4, un indicio de que la estratificación económica no sigue una única pendiente disciplinar a lo largo del ranking.

La dimensión organizacional aporta un componente adicional: la estructura de precios difiere entre conglomerados editoriales. Dentro del conjunto analizado en este estudio, Frontiers Media SA registra el APC promedio más alto (2 901,91 USD), seguida por Springer Nature (2 653,16 USD); un segundo bloque agrupa a Taylor & Francis, Wiley y SAGE ($\approx 2\,192$ – $2\,246$ USD), mientras MDPI (2 078,41 USD) y Elsevier (1 949,82 USD) quedan por debajo dentro del conjunto de grandes editoriales, y el grupo “Otros” se sitúa notablemente más bajo (861,19 USD). Este gradiente es consistente con la evidencia de concentración de ingresos por APC en grandes publishers. Butler et al. (2023) estimaron que cinco editoriales grandes capturaron 1,06 mil millones de USD en ingresos por APC entre 2015 y 2018, con 589,7 millones atribuidos a Springer Nature, lo que dimensiona la escala económica en la que opera el segmento superior del mercado y contextualiza las diferencias organizacionales observadas en los APC.

En este marco interpretativo, es necesario considerar algunas limitaciones del alcance analítico. Los tiempos examinados corresponden a valores promedio declarados por las revistas y no a intervalos medidos artículo por artículo, lo que puede ocultar variaciones internas amplias dentro de un mismo título (Runde 2021; Zabala et al. 2023). Asimismo, el universo se restringe a revistas activas indexadas simultáneamente en Scopus y registradas en DOAJ, por lo que los hallazgos describen un segmento específico del acceso abierto y no la totalidad del ecosistema OA. En particular, el análisis empírico de costos y su comparación estructural se limitó a revistas con APC declarado

mayor a cero, de modo que los patrones económicos identificados no son extrapolables a revistas sin cargos, incluidas aquellas que operan bajo esquemas diamante (Klebel y Ross-Hellauer 2023; Morrison et al. 2022). Finalmente, el uso exclusivo del SJR como indicador de posicionamiento no incorpora otras métricas de impacto que podrían matizar parcialmente la intensidad de las asociaciones observadas (Björk y Solomon 2015; Siler y Frenken 2020). En consecuencia, los resultados deben interpretarse dentro del recorte descriptivo y del alcance empírico aquí delimitado.

En suma, el contraste entre ambas dimensiones muestra que el acceso abierto indexado en Scopus y registrado en DOAJ no se organiza bajo un único mecanismo de estratificación. Los tiempos editoriales, con promedios concentrados y relación prácticamente nula con SJR, se comportan como una dimensión relativamente estable, con diferencias más visibles por disciplina y por publisher que por cuartil. En cambio, los APC exhiben una estructura escalonada por posicionamiento, con heterogeneidad creciente en los cuartiles inferiores y diferencias sistemáticas por área y conglomerado editorial. En conjunto, los resultados de este estudio aportan evidencia empírica para comprender que la apertura de acceso a la lectura convive con condiciones económicas heterogéneas para la publicación, y que la estructura de APC constituye un componente central de las dinámicas organizacionales y disciplinares del ecosistema Scopus–DOAJ (Borrego 2023; Grossmann y Brems 2021; Morrison et al. 2017; Morrison et al. 2022).

Referencias*

- Björk, Bo-Christer, y David Solomon. 2015. "Article Processing Charges in OA Journals: Relationship between Price and Quality." *Scientometrics* 103 (2): 373–85. doi: [10.1007/s11192-015-1556-z](https://doi.org/10.1007/s11192-015-1556-z).
- Borrego, Ángel. 2023. "Article Processing Charges for Open Access Journal Publishing: A Review." *Learned Publishing* 36 (3): 359–378. [10.1002/leap.1558](https://doi.org/10.1002/leap.1558).
- Budzinski, Oliver, Thomas Grebel, Jens Wolling, y Xijie Zhang. 2020. "Drivers of Article Processing Charges in Open Access." *Scientometrics* 124 (3): 2185–2206. [10.1007/s11192-020-03578-3](https://doi.org/10.1007/s11192-020-03578-3).
- Butler, Leigh-Ann, Lisa Matthias, Marc-André Simard, Philippe Mongeon, y Stefanie Haustein. 2023. "The Oligopoly's Shift to Open Access: How the Big Five Academic Publishers Profit from Article Processing Charges." *Quantitative Science Studies* 4 (4): 778–99. [10.1162/qss_a_00272](https://doi.org/10.1162/qss_a_00272).
- Chen, Tsung-An, Ming-Hwai Lin, Yu-Chun Chen, y Tzeng-Ji Chen. 2024. "The Time from Submission to Publication in Primary Health Care Journals: A Cross-Sectional Study." *Publications* 12 (2): 13. [10.3390/publications12020013](https://doi.org/10.3390/publications12020013).
- Costello, Emer. 2019. "Bronze, Free, or Fourrée: An Open Access Commentary." *Science Editing* 6 (1): 69–72. [10.6087/kcse.157](https://doi.org/10.6087/kcse.157).
- Demeter, Márton, y Ronina Istratii. 2020. "Scrutinising What Open Access Journals Mean for Global Inequalities." *Publishing Research Quarterly* 36 (4): 505–22. [10.1007/s12109-020-09771-9](https://doi.org/10.1007/s12109-020-09771-9).
- Ellingson, Mallory K., Xiaoting Shi, Joshua J. Skydel, Kate Nyhan, Richard Lehman, Joseph S. Ross, y Joshua D. Wallach. 2021. "Publishing at Any Cost: A Cross-Sectional Study of the Amount that Medical Researchers Spend on Open Access Publishing Each Year." *BMJ Open* 11 (2): e047107. [10.1136/bmjopen-2020-047107](https://doi.org/10.1136/bmjopen-2020-047107).
- Escobar Espinoza, Andrés. 2023. "Retos de la Publicación Científica de Acceso Abierto." *Panorama Económico* 31 (4): 290–92. [10.32997/pe-2023-4791](https://doi.org/10.32997/pe-2023-4791).
- Ezema, Ifeanyi Jonas. 2019. "Gold Route Open Access Journals in Engineering and Technology: Analysis of Research Impact and Implications for Scholarly Communication." *Issues in Science and Technology Librarianship* (43). [10.29173/istl43](https://doi.org/10.29173/istl43).
- Fallahhusein, Muhamed, R. B. Fernandes, M. Naga Raju, I. S. S. Ugli, G. S. Nijaguna, y R. Udayakumar. 2025. "Open Access Service Models: Impact on Information System Design." *Indian Journal of Information Sources and Services* 15 (2): 148–53. [10.51983/ijiss-2025.IJISS.15.2.20](https://doi.org/10.51983/ijiss-2025.IJISS.15.2.20).
- González-Sanabria, Juan S., Alicia Aparicio, Diego A. Burgos-Pineda, y Fabián D. Guiza-Pinzón. 2020. "Indexing and Summary Services, an Undervalued Opportunity to Improve the Visibility and Dissemination of Articles Published in Latin American Scientific Journals: DOAJ Case." *Información Tecnológica* 21 (5): 223–34. [10.4067/S0718-07642020000500223](https://doi.org/10.4067/S0718-07642020000500223).
- Gray, Russell J. 2020. "Sorry, We're Open: Golden Open-Access and Inequality in Non-Human Biological Sciences." *Scientometrics* 124 (2): 1663–75. [10.1007/s11192-020-03540-3](https://doi.org/10.1007/s11192-020-03540-3).

* Última fecha de consulta de los sitios web: 13 de enero de 2026.

- Grossmann, Alexander, y Björn Brembs. 2021. "Current Market Rates for Scholarly Publishing Services." *F1000Research* 10: 20. [10.12688/f1000research.27468.2](https://doi.org/10.12688/f1000research.27468.2).
- Horbach, Serge P. J. M. 2020. "Pandemic Publishing: Medical Journals Strongly Speed Up Their Publication Process for COVID-19." *Quantitative Science Studies* 1 (3): 1056–67. [10.1162/qss.a.00076](https://doi.org/10.1162/qss.a.00076).
- Johann, David, Jörg Neufeld, Kathrin Thomas, Justus Rathmann, y Heiko Rauhut. 2024. "The Impact of Researchers' Perceived Pressure on Their Publication Strategies." *Research Evaluation*. [10.1093/reseval/rvae011](https://doi.org/10.1093/reseval/rvae011).
- Jubran, Jubran H., Lea Scherschinski, Dimitri Benner, Marian T. Park, Emmajane G. Rhodenhisser, Sufyan Ibrahim, M. Maher Hulou, Rohin Singh, Katherine Karahalios, Visish M. Srinivasan, Christopher S. Graffeo, y Michael T. Lawton. 2023. "Publication Speed Across Neurosurgery Journals: A Bibliometric Analysis." *World Neurosurgery*. [10.1016/j.wneu.2022.12.008](https://doi.org/10.1016/j.wneu.2022.12.008).
- Khoo, Shaun Yon-Seng. 2019. "Article Processing Charge Hyperinflation and Price Insensitivity: An Open Access Sequel to the Serials Crisis." *LIBER Quarterly* 29 (1): 1–18. [10.18352/lq.10280](https://doi.org/10.18352/lq.10280).
- Klebel, Thomas, y Tony Ross-Hellauer. 2023. "The APC-Barrier and Its Effect on Stratification in Open Access Publishing." *Quantitative Science Studies* 4 (1): 22–43. [10.1162/qss.a.00245](https://doi.org/10.1162/qss.a.00245).
- Krauskopf, Erwin. 2021. "Article Processing Charge Expenditure in Chile: The Current Situation." *Learned Publishing* 34 (4): 637–46. [10.1002/leap.1413](https://doi.org/10.1002/leap.1413).
- Meagher, Kate. 2021. "Introduction: The Politics of Open Access — Decolonizing Research or Corporate Capture?" *Development and Change* 52 (2): 340–58. [10.1111/dech.12630](https://doi.org/10.1111/dech.12630).
- Miguel, Sandra, Nancy Diana Gómez, y Paola Carolina Bongiovani. 2012. "Acceso Abierto Real y Potencial a la Producción Científica de un País. El Caso Argentino." *El Profesional de la Información* 21 (2): 146–53. [10.3145/epi.2012.mar.04](https://doi.org/10.3145/epi.2012.mar.04).
- Mingers, John, y Loet Leydesdorff. 2015. "A Review of Theory and Practice in Scientometrics." *European Journal of Operational Research* 246 (1): 1–19. [10.1016/j.ejor.2015.04.002](https://doi.org/10.1016/j.ejor.2015.04.002).
- Morrison, Heather, Widlyne Brutus, Myriam Dumais-Desrosiers, Tanoh Laurent Kakou, Katherine Laprade, Salah Merhi, Arbia Ouerghi, Jihane Salhab, Victoria Volkanova, y Sara Wheatley. 2017. "Open Access Article Processing Charges (OA APC) Longitudinal Study 2016 Dataset." *Data* 2 (2): 13. [10.3390/data2020013](https://doi.org/10.3390/data2020013).
- Morrison, Heather, Luan Borges, Xuan Zhao, Tanoh Laurent Kakou, y Amit Nataraj Shanbhog. 2022. "Change and Growth in Open Access Journal Publishing and Charging Trends 2011–2021." *Journal of the Association for Information Science and Technology*. [10.1002/asi.24717](https://doi.org/10.1002/asi.24717).
- Nabyonga-Orem, Juliet, James A. Asamani, Thomas Nyirenda, y Seye Abimbola. 2020. "Article Processing Charges Are Stalling the Progress of African Researchers: A Call for Urgent Reforms." *BMJ Global Health* 5 (9): e003650. [10.1136/bmjgh-2020-003650](https://doi.org/10.1136/bmjgh-2020-003650).
- Okagbue, Hilary I., Jaime Teixeira da Silva, y Timothy Anake. 2020. "Exploring the Relationship between Journal Indexing and Article Processing Charges of Journals Published by MDPI, the Multidisciplinary Digital Publishing Institute." *European Science Editing*. [10.3897/ese.2020.e54523](https://doi.org/10.3897/ese.2020.e54523).

Olejniczak, Anthony J., y Molly J. Wilson. 2020. "Who's Writing Open Access (OA) Articles? Characteristics of OA Authors at Ph.D.-Granting Institutions in the United States." *Quantitative Science Studies* 1 (4): 1429–1450. [10.1162/qss_a_00091](https://doi.org/10.1162/qss_a_00091).

Pallares, César, Gabriel Vélez Cuartas, Alejandro Uribe-Tirado, Diego Restrepo, Jaider Ochoa, y Marcela Suárez. 2022. "Situación del Acceso Abierto y los Pagos por APC en Colombia. Un Modelo de Análisis Aplicable a Latinoamérica." *Revista Española de Documentación Científica* 45 (4): e342. [10.3989/redc.2022.4.1931](https://doi.org/10.3989/redc.2022.4.1931).

Perdomo, Bexi, y Oscar Alberto Morales. 2022. "Editorial Policies of Communications in Open Access Journals Indexed in Scopus." *Investigación Bibliotecológica* 36 (93): 135–51. [10.22201/iib.24488321xe.2022.93.58650](https://doi.org/10.22201/iib.24488321xe.2022.93.58650).

Perdomo, Bexi, César Moreno More, y Marco Flores. 2023. "Revistas de Arquitectura en Acceso Abierto en Scopus: Análisis de Políticas Editoriales." *Revista General de Información y Documentación* 33 (2): 411–430. [10.5209/RGID.85408](https://doi.org/10.5209/RGID.85408).

Piwozar, Heather, Jason Priem, Vincent Larivière, Juan Pablo Alperin, Lisa Matthias, Bree Norlander, Ashley Farley, Jevin West, y Stefanie Haustein. 2018. "The State of OA: A Large-Scale Analysis of the Prevalence and Impact of Open Access Articles." *PeerJ* 2018 (2). [10.7717/peerj.4375](https://doi.org/10.7717/peerj.4375).

Raudales-García, Elí Vicente, Wileidys Artigas, Benjamín Barón Velandia, Nicolás Sumba Nacipucha, y Jorge Cueva Estrada. 2024. "Procesos Editoriales en Revistas Científicas: Un Análisis Bibliométrico a Partir de Artículos en Acceso Abierto." *Biblios* (87): e008. [10.5195/biblios.2024.1179](https://doi.org/10.5195/biblios.2024.1179).

Reyes Rodríguez, Alixon, y Rodrigo Moraga Muñoz. 2020. "Criterios de Selección de una Revista Científica para Postular un Artículo: Breve Guía para no 'Quemar' un Paper." *Sophia* 16 (1): 93–109. [10.18634/sophiaj.16v.1i.977](https://doi.org/10.18634/sophiaj.16v.1i.977).

Rodrigues, Rosângela Schwarz, Ernest Abadal, y Breno Kricheldorf Hermes de Araújo. 2020. "Open Access Publishers: The New Players." *PLOS ONE* 15 (6): e0233432. [10.1371/journal.pone.0233432](https://doi.org/10.1371/journal.pone.0233432).

Runde, Brendan J. 2021. "Time to Publish? Turnaround Times, Acceptance Rates, and Impact Factors of Journals in Fisheries Science." *PLOS ONE* 16 (9): e0257841. [10.1371/journal.pone.0257841](https://doi.org/10.1371/journal.pone.0257841).

Sánchez-Tarragó, Nancy, Alejandro Caballero-Rivero, Piotr Trzesniak, Dania Deroy Domínguez, Raimundo Nonato Macedo dos Santos, y Juan Carlos Fernández-Molina. 2016. "Las Revistas Científicas en América Latina hacia el Camino del Acceso Abierto: Un Diagnóstico de Políticas y Estrategias Editoriales." *Transinformação* 28 (2): 159–72. [10.1590/2318-08892016000200003](https://doi.org/10.1590/2318-08892016000200003).

Schifini, Luiz Roberto, y Rosângela Schwarz Rodrigues. 2019. "Política de Avaliação de Periódicos nas Áreas de Medicina: Impactos sobre a Produção Editorial Brasileira." *Perspectivas em Ciência da Informação* 24 (4): 78–111. [10.1590/1981-5344/3745](https://doi.org/10.1590/1981-5344/3745).

Schönfelder, Nina. 2020. "Article Processing Charges: Mirroring the Citation Impact or Legacy of the Subscription-Based Model?" *Quantitative Science Studies* 1 (1): 6–27. [10.1162/qss_a_00015](https://doi.org/10.1162/qss_a_00015).

Segado-Boj, Francisco, Juan Jose Prieto-Gutiérrez, y Juan Martín-Quevedo. 2022. "Attitudes, Willingness, y Resources to Cover Article Publishing Charges." *Learned Publishing* 35 (4): 489–98. [10.1002/leap.1455](https://doi.org/10.1002/leap.1455).

- Shen, Cenyu, y Bo-Christer Björk. 2015. “‘Predatory’ Open Access: A Longitudinal Study of Article Volumes and Market Characteristics.” *BMC Medicine* 13 (1). [10.1186/s12916-015-0469-2](https://doi.org/10.1186/s12916-015-0469-2).
- Siler, Kyle, y Koen Frenken. 2020. “The Pricing of Open Access Journals.” *Quantitative Science Studies* 1 (1): 28–59. [10.1162/qss_a_00016](https://doi.org/10.1162/qss_a_00016).
- Smith, Audrey C., Leandra Merz, Jesse B. Borden, Chris K. Gulick, Akhil R. Kshirsagar, y Emilio M. Bruna. 2021. “Assessing the Effect of Article Processing Charges on the Geographic Diversity of Authors.” *Quantitative Science Studies* 2 (4): 1123–43. [10.1162/qss_a_00157](https://doi.org/10.1162/qss_a_00157).
- Solomon, David J., y Bo-Christer Björk. 2012. “A Study of Open Access Journals Using Article Processing Charges.” *Journal of the American Society for Information Science and Technology* 63 (8): 1485–95. [10.1002/asi.22673](https://doi.org/10.1002/asi.22673).
- Soto-Herrera, Diego Alejandro, Fernanda Beigel, y César Pallares. 2025. “Costos de Publicación en Acceso Abierto bajo el Modelo APC.” *Revista Interamericana de Bibliotecología* 48 (1). [10.17533/udea.rib.v48n1e357790](https://doi.org/10.17533/udea.rib.v48n1e357790).
- Tennant, Jonathan P., y Daniel R. Lomax. 2019. “An Overview of Open Access Publishing in Palaeontology.” *Palaeontology Electronica* 22 (2). [10.26879/968](https://doi.org/10.26879/968).
- Tort, Adriano B. L., Zé H. Targino, y Olavo B. Amaral. 2012. “Rising Publication Delays Inflate Journal Impact Factors.” *PLOS ONE* 7 (12): e53374. [10.1371/journal.pone.0053374](https://doi.org/10.1371/journal.pone.0053374).
- Triggle, Chris R., Ross MacDonald, David J. Triggle, y Donald Grierson. 2022. “Requiem for Impact Factors and High Publication Charges.” *Accountability in Research* 29 (3): 133–64. [10.1080/08989621.2021.1909481](https://doi.org/10.1080/08989621.2021.1909481).
- White, Richard Kenneth Alistair, Angel Angelo, Daniel Fitchett, Murray Fraser, Liz Hayes, Jessica Howie, Erica Richardson, y Bianca White. 2021. “Only Two out of Five Articles by New Zealand Researchers Are Free-to-Access: A Multiple API Study of Access, Citations, Cost of Article Processing Charges (APC), and the Potential to Increase the Proportion of Open Access.” *PeerJ* 9: e11417. [10.7717/peerj.11417](https://doi.org/10.7717/peerj.11417).
- Wilches-Visbal, Jorge Homero, y Midian Clara Castillo-Pedraza. 2022. “Indicadores Bibliométricos: Impacto y Altmetrics.” *MedUNAB* 25 (1): 5–8. [10.29375/01237047.4363](https://doi.org/10.29375/01237047.4363).
- Yuen, Jason, Samiul Muquit, y Peter C. Whitfield. 2018. “Correlation Between Cost of Publication and Journal Impact. Comprehensive Cross-Sectional Study of Exclusively Open-Access Surgical Journals.” *Journal of Surgical Research*. [10.1016/j.jsurg.2018.06.029](https://doi.org/10.1016/j.jsurg.2018.06.029).
- Zabala, Jon, Borja González-Albo, Ana García-García, Aurora Garrido-Domínguez, José Ignacio Vidal-Li, Luis R. Álvarez-Díez, Soledad Hernando-Tundidor, Yara Mostazo-Fernández, y Teresa Abejón. 2023. “Evaluation and Publication Delay in Ibero-American Scientific Journals.” *Learned Publishing* 36 (2): 205–16. [10.1002/leap.1497](https://doi.org/10.1002/leap.1497).
- Zimmer, Rafael Leal, Aline Castello Branco Mancuso, Ursula Matte, y Patricia Ashton-Prolla. 2023. “Analysis of the Interval between Submission and Publication in Genetics Journals.” *PLOS ONE* 18 (5): e0284866. [10.1371/journal.pone.0284866](https://doi.org/10.1371/journal.pone.0284866).
- Zukerfeld, Mariano, Carolina Unzurrunzaga, y Carolina Monti. 2023. “Ranking, Reconocimiento y Cargos por Publicación (APC).” *Palabra Clave (La Plata)* 12 (2): e183. [10.24215/18539912e183](https://doi.org/10.24215/18539912e183).

Community-Driven Digitization of Ancient Manuscripts: Communication Strategies and Organizational Practices in Wikisource Indonesia*

Fathul Qorib^(a)
Hafidlatul Fauzuna^(b)
Qurrotul Aini^(c)
Noor Khalida Magfirah^(d)
Samia Elviria^(e)

a) University of Tribhuwana Tunggaladewi, <https://orcid.org/0000-0001-8081-3252>
b) State Islamic University of Madura, <https://orcid.org/0009-0001-0478-2551>
c) State Islamic University of Madura, <https://orcid.org/0009-0009-9431-8012>
d) University of Jambi, <https://orcid.org/0009-0004-8714-3238>
e) University of Nurdin Hamzah, <https://orcid.org/0009-0009-1681-8441>

Contact: Fathul Qorib, fathul.qorib@unitri.ac.id; Hafidlatul Fauzuna, fauzunafieda@iainmadura.ac.id;
Qurrotul Aini, qurrotula823@gmail.com; Noor Khalida Magfirah, noorkhalidamagfirah@unja.ac.id;
Samia Elviria, samiaeelviria@unh.ac.id

Received: 13 October 2025; **Accepted:** 20 March 2026; **First Published:** 15 May 2026

ABSTRACT

This study examines the role of the Wikisource Indonesia Community, a volunteer-based organization, in digitizing ancient manuscripts. It explores the communication strategies and organizational practices that help the community manage this project. Using a qualitative approach, the research involved interviewing key members, including the founder, manuscript experts, and volunteers who are directly involved in the digitization work. Data was collected over seven months through in-depth interviews and participant observation, where the researcher attended meetings, recorded discussions, and observed digitization activities. The data was then analyzed to identify key themes and cross-checked for accuracy. The findings show that effective communication, both within the community and with external partners, is crucial for smooth coordination. The community stays connected using WhatsApp groups and Zoom meetings, even though members are spread out geographically. However, there were common challenges like miscommunication, technical issues, and leadership gaps. Despite these hurdles, a strong sense of identity and a shared commitment to preserving Indonesia's cultural heritage kept volunteers motivated. Regular updates and transparent decision-making also helped the community overcome these challenges. This study sheds light on how volunteer-driven projects can successfully preserve cultural heritage and highlights the importance of shared purpose and open communication in keeping volunteers engaged and motivated.

KEYWORDS

Community-based preservation; Digitisation; Organizational communication; Wikisource.

* Fathul Qorib was responsible for the overall research design, conceptualization, and lead writing of the manuscript. Hafidlatul Fauzuna and Noor Khalida Magfirah conducted the data collection, including the interviews with key members and participant observation of the digitization activities. Qurrotul Aini and Samia Elviria contributed to the literature review on organizational communication and the final editing of the manuscript. All authors have read and approved the final version of the paper.

© 2026, The Author(s). This is an open access article, free of all copyright, that anyone can freely read, download, copy, distribute, print, search, or link to the full texts or use them for any other lawful purpose. This article is made available under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. JLIS.it is a journal of the SAGAS Department, University of Florence, Italy, published by EUM, Edizioni Università di Macerata, Italy, and FUP, Firenze University Press, Italy.

1. Introduction

Digitization of ancient manuscripts is important for preserving culture and knowledge, enabling manuscripts vulnerable to physical damage to be preserved and accessible to future generations (Fakhriati et al. 2022). Digitization expands public access to the information and knowledge in ancient manuscripts. In addition, digitization protects manuscripts from physical damage due to age and environmental conditions. The digitization of ancient manuscripts is growing rapidly in Indonesia. For example, the Aceh Library and Archives Agency has digitized 132 ancient manuscripts on various topics. The Wikisource Loves Manuscripts project has digitized more than 20,000 pages of manuscripts from Bali, Java and Sumatra. In Yogyakarta, WILMA digitized 33 manuscripts with over 8,000 pages (VOA Indonesia 2023).

In Indonesia, Wikisource plays an important role in the digitization of ancient manuscripts. Wikisource is a free library that includes source texts, translations, historical texts, and biographies, with approximately 3,955 publicly accessible texts (Wikisource Bahasa Indonesia 2024). Wikimedia and Wikisource digitize many manuscripts, including source texts, translations, historical texts, and author bibliographies. The Wikisource Loves Manuscripts (WLM) project has digitized Quranic manuscripts, gending manuscripts, and other Javanese manuscripts (VOA Indonesia 2023). By 2023, the digitization of ancient manuscripts continued to grow and involved projects across organizations and communities such as WLM, the Centre for the Study of Islam and Society (PPIM) and UNESCO, digitizing thousands of pages of manuscripts (Wikimedia Foundation 2023).

Data from the Grand Design of Nusantara Manuscript Management records 121,668 titles of Nusantara manuscripts scattered around the world, with 12,358 titles of ancient manuscripts in 23 languages in the archipelago that have been successfully digitized. This digitization effort includes digital conversion and involves physical treatment and systematic cataloguing to ensure widespread accessibility of the manuscripts (VOA Indonesia 2023). However, significant challenges still arise behind these advancements, particularly related to the lack of public interest and awareness that has led to low literacy of ancient manuscripts (Fischer et al. 2010; Sidhartani 2020). Despite the availability of digital technology, the adoption and use of digital platforms are still limited, which hinders the full utilization of digitization's potential (Arinta and Widyastuti 2024; Isnah et al. 2022). In addition, low information literacy further exacerbates the barriers to interacting with these valuable historical texts (Hadiati et al. 2024).

Several previous studies emphasize manuscript digitization's important role in protecting valuable cultural information from environmental damage and neglect. Digitization typically involves scanning, storing, and making manuscripts accessible online, thus expanding their reach to researchers and the general public (Fakhriati et al. 2022; Kuswara 2018). Libraries and museums have been key players in this effort, with institutions such as Reksa Pustaka Mangkunegaran and Museum Radya Pustaka Solo actively digitizing their collections (Marintan et al. 2024; Sadono et al. 2025). We can also see from the (Prameswari et al. 2024) works that digitization projects have grown globally, as reflected in the bibliometric analysis.

Digitization goes beyond protecting fragile documents from damage and loss; it also enables new ways of interacting with cultural materials, which supports sustainable heritage management (Abdoh 2025; Fanea-Ivanovici 2018; Kumar and Shah 2018). However, this process is not easy. It re-

quires the involvement of manuscript owners and communities to maintain cultural sustainability and knowledge transfer (Mengping 2023; Wibowo et al. 2024). Technological advances in image optimization and capture methods are improving how digitized manuscripts can be accessed and searched. However, challenges remain in preserving these manuscripts' original context and intangible cultural values (Chisita et al. 2021; Suryana et al. 2024). Successful digitization projects often rely on collaborative approaches, disaster preparedness, and effective information systems, underscoring the work's complex social and technological interactions (Che Hasssan et al. 2024; Nistane and Mishra 2024). However, the unfortunate thing about ancient manuscripts in Indonesia is that they are still stored in European institutions, so Indonesia needs to accelerate comprehensive digitization initiatives (Aimah 2021).

Despite these advances, important gaps remain, especially regarding the social dynamics and organizational processes that support digitization efforts, especially within volunteer-based communities like Wikisource Indonesia. Typically, digitization efforts succeed when the government supports them through official institutions and substantial funding (Wibowo 2023). On the other hand, community-based digitization is important to directly involve the community, which is not working as part of a job and earning money but voluntarily out of love for manuscript preservation. While many studies have focused on the technical aspects and institutional framework of digitization, little is known about how communication strategies among dispersed volunteers affect the efficiency and quality of manuscript digitization. In addition, while Indonesia's low public interest and digital literacy have been recognized as barriers to adoption, there is limited understanding of how community-driven initiatives can leverage collaborative communication and management practices to overcome these challenges.

This gap suggests the need to explore the human and organizational factors contributing to successful digitization, offering new insights to improve technology implementation and cultural engagement. Therefore, this research investigates the communication strategies and organizational practices the Wikisource Indonesia Community uses in digitizing ancient manuscripts, aiming to deepen the understanding of sustainable cultural preservation in the digital age. Specifically, this research examines how communication keeps volunteers connected and motivated during the digitization process and how day-to-day organizational practices overcome challenges to maintain smooth operations. In addition, it explores how these approaches foster cultural interest and public awareness of cultural heritage preservation in today's digital world. By focusing on these interconnected aspects, this research seeks to provide a comprehensive picture of what drives truly sustainable and meaningful digital preservation.

2. Literature Review

2.1 Digitisation of Old Manuscripts

Digitization of ancient manuscripts involves converting physical artifacts into digital formats through structured methodologies to preserve cultural heritage and increase accessibility (Sančanin and Penjišević 2023). This transformation reshapes how knowledge and culture are perceived and engaged with, giving rise to the concept of digital heritage, which supports innovative research and greater public interaction with historical documents (He et al. 2017). The initial stages typical-

ly include the selection of manuscripts for digitization, followed by high-quality scanning to produce digital images of the originals (Sporleder 2010). These images are then processed with optical character recognition (OCR) to convert into machine-readable texts, enhancing their searchability and usability in digital environments (Colavizza et al. 2019).

Following OCR, detailed metadata is created to capture visual and contextual information, ensuring effective categorization and management of digitized manuscripts (Pandey and Kumar 2020). Quality assurance through readability testing and validation by the digitization team helps maintain the integrity and accuracy of the texts. Once validated, these manuscripts are published on publicly accessible digital platforms, making them widely available for research and cultural engagement (Timofte and Van Gool 2013). Altogether, these interconnected steps represent a comprehensive process that balances preservation with accessibility, facilitating sustainable cultural heritage management in the digital age.

Significantly, digitization offers many advantages, including increased accessibility, but also presents challenges for cultural heritage institutions. According to Razali et al., the growth of community-driven initiatives emphasizes the importance of creating digital spaces that reflect the collective knowledge of local cultures (Razali et al. 2009). Such projects increase awareness and appreciation of cultural heritage and contribute to sustainable community engagement through technology-based platforms (Bruton 2011). Economic concerns remain as many digitization projects lack adequate funding, highlighting the need for a stable policy and cooperation framework to encourage digital preservation (Ahmad and Rafiq 2023; Singh 2012).

The digital preservation landscape continues to evolve with technological advances such as 3D modelling and virtual reality, which offer new ways to represent cultural heritage (Zou et al. 2024). The role of metadata systems becomes apparent as they ensure the long-term sustainability and accessibility of digitized manuscripts by organizing information for efficient searching (Hu et al. 2017; Masenya 2021). Digitization, therefore, acts as a protective measure for manuscripts and facilitates lasting engagement with the cultural narratives they contain, bridging the gap between past and present for future users.

2.2 Organizational Communications

Effective organizational communication is essential for coordination and collaboration at every stage of the digitization process. Organizational communication refers to the flow of information between different organizational levels and functions, including formal and informal channels (Bell and Leonard 2016). This communication is heavily influenced by the organization's culture, which shapes how messages are sent and received. A strong culture characterized by openness and transparency fosters an environment where employees freely share information and collaborate, which is an important factor, especially in complex digitization projects (Kocak and Pawlowski 2023; Laaksonen and Porttikivi 2021). As Kocak and Pawlowski emphasize, digitization requires collaborative engagement between different departments, which requires organizational structures to evolve and support new communication and collaboration methods.

There are two main types of organizational communication that are particularly relevant in digitalization: internal communication and strategic communication. Internal communication focuses

on interactions among employees and departments, ensuring that everyone is well-informed about digitization goals and progress, which helps drive engagement and inclusiveness (Badham et al. 2022). Meanwhile, strategic communication aligns overall organizational goals with digitization efforts, providing a framework to communicate effectively with all stakeholders and reinforcing the importance of the project (Badham et al. 2022; Mohamad et al. 2018). The increasing reliance on digital communication tools greatly affects organizational culture and communication effectiveness, requiring organizations to adopt innovative strategies to sustain meaningful dialogue and collaboration (Poltoratska et al. 2024).

Collaboration in digitalization also relies heavily on digital tools that facilitate communication and project management. Platforms like Slack, Trello, and Asana help overcome geographical barriers and enable real-time interactions among team members, streamlining workflows and increasing social capital through knowledge sharing and relationship building (Olaniyi et al. 2024). To navigate the complexities of digital communication, clear policies and guidelines must be established to organize interactions and ensure smooth project execution (Banghart et al. 2018; Knudsen and Nielsen 2019). Overcoming barriers such as organizational silos and resistance to change requires committed leadership to foster a culture that embraces innovation and values open communication (Abhari et al. 2021; Aris and Hassan 2023). Failure to integrate effective communication strategies can significantly reduce the likelihood of achieving digitization goals, highlighting the critical role of communication in this transformative process (Lock and Ludolph 2019).

3. Research Methods

This research uses a qualitative approach to explore how the Wikisource Indonesia Community communicates and organizes themselves while digitizing ancient manuscripts (Cresswell 2007). Primary data was collected through in-depth interviews with key informants, including the founder, manuscript resource persons, and community volunteers directly involved in the digitization activities. These informants were selected using a purposive sampling technique, ensuring that those interviewed were most competent and relevant to the research topic (Denzin and Lincoln 2009). The interviews provided detailed insights into the communication strategies and challenges faced by the community (Kallio et al. 2016; Rosenthal 2016).

Additionally, participatory observation was conducted over 7 months, during which the researcher actively engaged in community meetings, observed the digitization process, and documented ongoing activities. This extended observation helped the researcher understand the decision-making processes, communication flow, and collaboration dynamics within the community. Secondary data was obtained from meeting notes, event reports, and workshop summaries to triangulate the findings and ensure validity (Flick 2018). Table 1 below shows the list of informants and their roles:

Table 1. Research informants

No	Name Kode	Role
1	KAL	Founder of Wikisource Indonesia
2	DH	Wikipedia Indonesia Community Staff
3	EDN	Wikisource Community Contact Person 2023-2024
4	ARA	Volunteer 1
5	QW	Volunteer 2
6	AM	Wikisource Workshop Alumni 2023
7	SH	Wikisource Competition Participants 2024

Data analysis was performed by reviewing both primary and secondary data to gain a comprehensive understanding of the community's practices and challenges during digitization. The researcher utilized thematic analysis to examine the data, which allowed for a deeper exploration of communication patterns and insights into the decision-making processes within the community. The process began with data reduction, where the researcher sorted and condensed information to highlight key themes and relevant insights (Miles et al. 2020). Triangulation was applied to validate the findings by cross-referencing interview data with secondary documents, ensuring robustness and reliability (Flick 2018). Additionally, member checking was done by sharing interpretations with participants for feedback, and peer debriefing involved consulting with colleagues for objective critique to strengthen the trustworthiness and rigor of the findings (Elo et al. 2014).

4. Result

Digitization requires efficient workflows to manage elements in digitizing natural history collections and cultural heritage materials (Haston et al. 2012). High-yield workflows and advanced scanning tools ensure timely and accurate digitization (Hidalga et al. 2020; Willemse et al. 2020). Wikisource Indonesia is key to digitizing Indonesian literature and manuscripts, preserving documents, and making them accessible. Maintaining relevant metadata and contextual information ensures that digitized data can be correctly accessed and understood by future users (Mittal et al. 2022).

4.1 Dynamic Communication

In the study of organizational communication theory, various patterns regulate information flow within structures (Musheke and Phiri 2021). The chain pattern depicts a linear hierarchy; the Y pattern distributes from one source to branches, the wheel relies on a central individual, the circle facilitates adjacent exchanges, the all-channel allows direct interactions, and the network supports complex pathways (Joseph and Gaba 2020). In the Wikisource Indonesia community, the communication pattern resembles the all-channel pattern but with more flexibility. Members interact directly and transparently through WhatsApp groups, fostering a collaborative ecosystem based

on trust and openness. This platform allows dynamic exchanges where every voice is valued, and contributions are recognized.

“Since its inception, the Wikisource community has operated online with members from different regions collaborating on the addition and proof reading of texts. We hold monthly meetings to decide which manuscripts to digitise and initiate a wiki fund program to find and nominate manuscripts.” (Interview with EDN)

The Wikisource Community adopts a structured yet adaptive approach to manuscript digitization, enabling geographically diverse members to collaborate online. Monthly meetings serve as forums for text selection, while initiatives like the “wiki fund,” which provides cash allowances, credit replacements, activity consumption, and souvenirs, further support member participation. This flexible and supportive communication model ensures effective resource management and active involvement from all members.



Figure 1. Flowchart of dynamic communication patterns in Wikisource Indonesia

Communication in these organizations occurs dynamically through various communication technologies such as WhatsApp, Zoom, and Google Meet. Instant messaging groups allow members in remote locations to connect and collaborate effectively. In addition, video conferencing platforms such as Zoom and Google Meet are used for online meetings, which support in-depth discussions and coordination between members from different locations so that the flow of information remains smooth and volunteers can actively engage. These diverse communication technologies also create flexibility and increase efficiency in collaboration (Dornelles et al. 2022). WhatsApp groups facilitate direct and equal communication, while Zoom and Google Meet provide spaces for more structured virtual meetings. This combination strengthens members’ integration by ensuring they can participate in meetings and discussions without geographical restrictions.

The advantage of dynamic communication over other organizational communication models is its ability to effectively overcome distance and time constraints. Modern communication technologies enable real-time interaction and collaboration without the need to meet in person, which is especially important for organizations with members spread across multiple locations. This

approach speeds up the flow of information. It increases engagement and responsiveness, making it more adaptive and efficient than traditional communication models, which are often hampered by physical and temporal limitations.

“Although KWI (Wikisource Indonesia Community) has a number of social media such as Gmail, Facebook, X, Instagram, and Telegram, none has been able to replace the role of WhatsApp groups as the main medium of internal communication” (Interview with EDN)

WhatsApp is still the leading choice for internal communication in the Wikisource Indonesia Community despite the availability of various other digital platforms such as Gmail, Facebook, X, Instagram, and Telegram. This is due to WhatsApp’s ability to facilitate quick, informal, and accessible communication for almost all community members. WhatsApp groups provide a simple yet effective conversation space, allowing daily coordination, light discussions, and instant responses without technical barriers. With high user familiarity, WhatsApp creates an essential sense of closeness and connectedness within the volunteer community. However, if not accompanied by adequate structure, this highly fluid communication risks diluting the focus on task completion.

While the fluid communication at Wikimedia Indonesia supports good connectedness, it often causes members to lack focus on task completion. Organizations need to set clear goals and responsibilities and use a project management system to remedy this. Regular meetings and progress reports can increase accountability and ensure members stay on track (Oghuvbu et al. 2022). In addition, structured communication tools help to collate information better. Wikimedia Indonesia can maintain effective communication with this approach while achieving optimal work targets. Appreciation for achieving targets shows promising results from members’ collaboration and dedication.

4.2 The Manuscript Digitisation Process

The process of digitising ancient manuscripts on Wikimedia, especially within the scope of the Wikisource Indonesia community, is a structured and systematic collaborative activity. The initial stage of this process begins with scanning the selected manuscripts, which is then followed by data collection and categorisation of the manuscripts. Each scanned manuscript is uploaded to the Wikisource platform, where it is ready to be read by community members. The monthly community meeting is crucial in determining which manuscripts will be prioritised for read-testing. During the meeting, community members vote to select the manuscripts targeted for digitisation that month.

Once the text has been selected, community members jointly conduct a reading test to ensure the quality of the digitised text. This reading test process not only verifies the accuracy of the text but also ensures that it is free of typographical errors and complies with the standards set by Wikisource. This activity is reinforced by read-testing competitions, workshops, and other activities designed to increase participation and the quality of digitisation. Each member is expected to make at least two edits per week, with the contact person actively providing regular reminders and recaps of edits to ensure member engagement and project progress.

“We support manuscript conservation by proofreading and digitising ancient, copyright-free manuscripts on Wikisource, making them accessible to all. We also hold competitions, online meetings, and reading programs to speed up digitisation.” (Interview with KAL).

“The Wikisource community operates online, editing and proofreading manuscripts. We discuss and vote on monthly digitisation targets in meetings and have a wiki fund activity where participants find manuscripts to digitise.” (Interview with ARA).

As such, the manuscript digitisation process at Wikisource Indonesia reflects an efficient cooperation model where digital technology and open communication are key to preserving Indonesia’s literary heritage. Through community members’ hard work and dedication, Indonesia’s ancient manuscripts have been saved from physical damage and made more accessible to the broader community, enriching the nation’s treasure of knowledge and culture.



Figure 2. Diagram of the manuscript digitisation process

The chart above illustrates the flow of the manuscript digitisation process in the Wikisource Indonesia Community. The process begins with selecting manuscripts to be digitised; then, the manuscripts are scanned using a special scanner. The scan results are recorded and categorised based on relevant information. After that, the manuscript is read-tested and validated by the team or community involved. The final stage is the publishing and announcement of the digitisation results through a digital platform that the public can access. Each stage requires good coordination and collaboration to achieve optimal results.

Manuscript digitisation by the Wikisource Indonesia Community uses a collaborative and participatory approach. The community uses digital technology to scan, record, and categorise copyright-free ancient manuscripts. Community members conduct the reading test and validation process spread across different regions, allowing for broad and diverse engagement. WhatsApp groups and monthly online meetings facilitate communication and coordination between community members. With this approach, the Wikisource Indonesia Community accelerated the digitisation process and ensured that the public could access and utilise ancient manuscripts in digital form.

4.3 Challenges of Manuscript Digitisation

The challenges and obstacles in the management of manuscript digitisation by Wikimedia Indonesia are diverse and complex, covering technical, structural, and communication aspects between members. Technical barriers are one of the main challenges, where accounts are often difficult to log in to, and there are technical differences between one volunteer and another (Pandey and Kumar 2020). Added to this is the lack of technical knowledge of the contact person, who sometimes doesn't realise or know how to solve problems faced by members.

“Communication in the community is sometimes hindered by overlaps between the roles of Community Staff and GLAM Staff, leading to miscommunication. On the other hand, strong member-to-member communication has been supported by a consistent contact person over time. The main challenge lies in exchanging opinions and the varied response times due to members' activities and Indonesia's time zone differences,” (Interview with DH)

Communication within Wikimedia Indonesia faces particular challenges, particularly in coordinating roles between the Community and GLAM (Gallery, Library, Archive, and Museum) staff. While platforms such as WhatsApp and the existence of long-standing contact persons support member interaction, overlapping responsibilities and unclear divisions of tasks often create friction—especially when activities require collaboration across structural boundaries. This internal ambiguity becomes more pronounced in digitization projects, where differing organizational hierarchies and job descriptions between community and GLAM staff contribute to misinterpretations and delays. The situation is further complicated by Indonesia's vast geography and spread of time zones, which hinder synchronous communication and slow collective responses. As volunteers operate across islands with varying daily routines, aligning efforts in real-time is a constant coordination challenge.

The delivery of messages through contact persons, while considered good enough, also faces obstacles. The speed of response often depends on the activities of individual members and the time difference in Indonesia. This poses a challenge in ensuring that information is received and understood correctly by all members. Direct feedback from community members is very helpful in facilitating the communication process. However, challenges arise when the feedback is not responded to quickly or is not implemented properly. This can cause frustration among volunteers and hamper the digitisation workflow.

Message evaluation is crucial to avoid misunderstandings, but monthly evaluation meetings often lack follow-up, leading to repeated mistakes. Introductory media like WhatsApp, Instagram, X, and Telegram help disseminate information, but not all members are active on all platforms, causing communication gaps. While Wikimedia Indonesia's funds, allowances, and souvenirs motivate volunteers, maintaining long-term enthusiasm is challenging. To ensure smooth digitisation, the community must increase sensitivity to miscommunication and improve the communication flow, requiring joint efforts from all members.

“To reduce communication barriers, we have a strategy that we members enter the community through invitation, so when you have participated in wikisource activities and diligently edited at home, you will receive an email invitation from the contact person to join the community. So it is certain that the members

are consistent and not toxic. For the language itself, we use Indonesian, both in the WhatsApp group and during meetings.” (Interview with ARA).

To strengthen this communication ecosystem, Wikimedia Indonesia could consider developing a unified and inclusive communication framework that combines informal channels with structured documentation and more precise follow-up mechanisms. As mentioned by the informant above, an invitation-based membership system helps ensure that community members are consistent, active, and non-toxic, creating a more focused collaborative environment. Using Indonesian in WhatsApp groups and meetings also reduces potential language barriers. However, to prevent communication gaps and ensure that feedback is acted upon effectively, the community must build shared tools—such as collaborative workspaces, shared calendars, and task-tracking systems—that support transparency and sustainability. These additions can reduce over-reliance on individual contact persons and encourage a more responsive and coordinated digitalization process.

5. Discussion

The study results show that organizational communication in the Wikisource Indonesia Community plays a role beyond simply exchanging information; it is a social mechanism that maintains motivation and strengthens trust among geographically dispersed members. Democratic selection of contact persons in community meetings creates a sense of ownership and collective responsibility critical to maintaining volunteer engagement. This open and transparent organizational culture supports a smooth flow of information and inclusive decision-making, allowing the community to adapt to challenges arising during digitalization. This emphasizes that effective communication in volunteer-based communities is formal, flexible, and oriented towards close interpersonal relationships (Reed and Hagen 2025). The roles of internal and strategic communication in this community complement each other to maintain smooth operations and build a shared vision. Internal communication via WhatsApp and Telegram ensures that all members understand the goals and progress, reducing the potential for miscommunication and increasing a sense of inclusivity (Estrada Molina 2022). Meanwhile, through meetings and official announcements, strategic communication confirms long-term goals and strengthens the commitment of all members to the digitalization project. This two-way communication dynamic balances the need for technical coordination and the strengthening of shared values so the community can function effectively even though its members are widely dispersed.

Using digital collaboration tools such as WhatsApp and Telegram has proven highly effective in overcoming geographical barriers and time zone differences. These tools enable real-time interactions that accelerate workflow, fostering social capital through knowledge exchange and interpersonal relationships. Interview data reveal that these technologies do more than facilitate technical tasks; they also strengthen social bonds among members, which is crucial in volunteer-based communities. This finding supports the notion that the success of digitization initiatives heavily relies on integrating technology with a resilient and adaptive social culture (Miceli et al. 2021). Furthermore, these platforms’ ease of access and familiarity contribute to sustained engagement, allowing members to collaborate flexibly despite physical separation.

Although communication within the community functions effectively overall, several challenges remain, including miscommunication between Wikimedia and GLAM staff, delayed responses, and limitations in the leadership skills of coordinators. These obstacles highlight the need for targeted communication training and conflict management development to enhance community leadership capacities. A deep understanding of social dynamics and technical constraints is essential to ensure that the digitization process proceeds smoothly without significant disruptions. Proactive efforts to cultivate interpersonal skills and formal communication competencies are necessary for the community's resilience and growth amid the complex environment in which they operate (Hobgood et al. 2002; Zamiri and Esmaeili 2024). Addressing these gaps can substantially improve decision-making efficiency and overall project outcomes.

The aspect of community identity also significantly contributes to communication effectiveness. The philosophy and symbolism embodied in the community's logo reflect shared values, strengthening a sense of togetherness and collective purpose. This shared vision motivates members to contribute consistently despite operational and geographical challenges. Effective communication within this community transcends mere functionality; it also fosters emotional bonds that enhance social cohesion and reinforce volunteers' commitment to digital cultural preservation. Moreover, the strong collective identity stabilizes during periods of uncertainty, enabling the community to navigate challenges collaboratively. Such identity-driven communication creates a resilient social fabric that sustains long-term engagement and fosters a culture of mutual support.

From a management perspective, this study confirms that the success of community-based digitization requires a combination of effective communication, collaborative technological support, and leadership that fosters a culture of innovation and openness. Organizations seeking to replicate this model must prioritize communication training and develop inclusive digital systems to ensure all stakeholders can fully participate and contribute to heritage preservation efforts (Chisita et al. 2021; Micle et al. 2023). Furthermore, cultivating leadership that embraces adaptability and encourages continuous learning is essential to navigating the evolving challenges of digital transformation. Institutional commitment to these elements is crucial in building sustainable frameworks empowering communities to manage cultural assets digitally.

6. Conclusion

Overall, organizational communication within the Wikisource Indonesia Community is a strategic element that bridges people, technology, and culture. This communication facilitates information exchange and nurtures a sense of shared purpose and collective identity crucial for volunteer engagement. The inclusive, adaptive, and participatory communication approach becomes essential in overcoming challenges and ensuring the sustainability of cultural heritage preservation in digital form. By fostering open dialogue and continuous feedback, the community can respond effectively to evolving needs and maintain cohesion despite geographical dispersion. Thus, this study reinforces that digitizing ancient manuscripts is not merely a technical transformation but a social one, requiring thoughtful communication management focused on community building. This study examines the pivotal role of the Indonesian Wikisource community in preserving national literary heritage through the digitization of ancient manuscripts. The collaborative efforts of community members demonstrate how digital technology, combined with traditional preservation

methods, can create a widely accessible digital archive that maintains the integrity of the original texts. This inclusive and participatory approach improves the quality and efficiency of digitization and strengthens community social ties, thereby fostering meaningful synergies between cultural preservation and sustainable community development. These findings underscore the importance of a well-organized community framework in addressing manuscript preservation challenges and promoting open access to cultural knowledge.

However, this study has limitations, particularly its focus on a single community, which may limit the generalizability of its findings to other social and technological contexts. The reliance on interviews and participant observation as primary data collection methods may not fully capture the complexity of communication dynamics and challenges faced during digitization. Moving forward, there is an urgent need for more comprehensive studies that build on these findings. Future research could involve comparative analysis across multiple digitizing communities, using mixed methods such as ethnographic observation and social network analysis to understand communication practices and their impacts better. Such work will help identify best practices tailored to diverse contexts, particularly for local communities seeking to preserve their cultural heritage through digital means.

References*

- Abdoh, Esra. 2025. "Utilizing Modern Technology for the Preservation of Ancient Manuscripts and Rare Books: The Digitization Project at King Abdulaziz Complex for Endowment Libraries as a Model." *Restaurator. International Journal for the Preservation of Library and Archival Material* 46 (1): 35–58. <https://doi.org/10.1515/res-2024-0016>.
- Abhari, Kaveh, Chloe Ostroff, Bailey Barcellos, and David N. Williams. 2021. *Co-Governance in Digital Transformation Initiatives: The Roles of Digital Culture and Employee Experience*. <https://doi.org/10.24251/hicss.2021.704>.
- Ahmad, Rafiq, and Muhammad Rafiq. 2023. "Preserving Cultural Heritage: Digital Preservation in Small Community Libraries." *Library Hi Tech News* 41 (3): 17–8. <https://doi.org/10.1108/lhtn-07-2023-0113>.
- Aimah, Rhomayda Alfa. 2021. "Koleksi Manuskrip Indonesia Dalam Katalog Online Eropa Dan Urgensi Digitalisasi." *Dinamika Penelitian: Media Komunikasi Penelitian Sosial Keagamaan* 21 (02): 220–45. <https://doi.org/10.21274/dinamika.2021.21.02.220-245>.
- Arinta, Yusvita Nena, and Emy Widyastuti. 2024. "Accelerating Digital Banking Transformation Through User Acceptance: A Study on the Millennial Cultural." *Maksimum* 14 (1): 74. <https://doi.org/10.26714/mki.14.1.2024.74-87>.
- Aris, Ain Z. Z., and Azlini Hassan. 2023. "Enhancing Organizational Commitment Through Effective Computer Communication Practices: A Proposal for Malaysian Organizations." *E3s Web of Conferences* 440: 04011. <https://doi.org/10.1051/e3sconf/202344004011>.
- Badham, Mark, Vilma Luoma-aho, Chiara Valentini, and Laura Lumimaa. 2022. "Digital Strategic Communication Through Digital Media-Arenas." In *Research Handbook on Strategic Communication*, edited by Jesper Falkheimer and Mats Heide, 416–30. <https://doi.org/10.4337/9781800379893.00035>.
- Banghart, Scott, Michael Etter, and Cynthia Stohl. 2018. "Organizational Boundary Regulation Through Social Media Policies." *Management Communication Quarterly* 32 (3): 337–73. <https://doi.org/10.1177/0893318918766405>.
- Bell, Emma, and Pauline Leonard. 2016. "Digital Organizational Storytelling on YouTube: Constructing Plausibility Through Network Protocols of Amateurism, Affinity, and Authenticity." *Journal of Management Inquiry* 27 (3): 339–51. <https://doi.org/10.1177/1056492616660765>.
- Bruton, Dean. 2011. "Theorizing Digital Cultural Heritage: A Critical Discourse." *Information Communication & Society* 14 (7): 1077–78. <https://doi.org/10.1080/1369118x.2010.542826>.
- Che Hasssan, Nor Hasni, Wan Satirah Wan Mohd Saman, and Norshila Shaifuddin. 2024. "Evaluating the Scholarly Trends on Issues, Strategies and Sustainable Development of Digitisation of Documentary Heritage: A Thematic Review." *Environment-Behaviour Proceedings Journal* 9 (SI18): 247–54. <https://doi.org/10.21834/e-bpj.v9isi18.5486>.

* Last date of consultation of the websites: 20 April 2026.

Chisita, Takaingenhamo Collence, Oluwole O. Durodolu, and Joseph Ngoaketsi. 2021. "Evaluating the Processes and Procedure of Digitalization Workflow." In *Digital Libraries - Advancing Open Science*, edited by Sadia Vancauwenbergh. <https://doi.org/10.5772/intechopen.96851>.

Colavizza, Giovanni, Maud Ehrmann, and Fabio Bortoluzzi. 2019. "Index-Driven Digitization and Indexation of Historical Archives." *Frontiers in Digital Humanities* 6. <https://doi.org/10.3389/fdigh.2019.00004>.

Cresswell, John C. 2007. *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*. London: SAGE Publication, Inc.

Denzin, Norman K, and Yvonna S Lincoln. 2009. *Handbook of Qualitative Research*. Yogyakarta: Pustaka Pelajar.

Dornelles, Jéssica De Assis, Néstor F. Ayala, and Alejandro G. Frank. 2022. "Smart Working in Industry 4.0: How Digital Technologies Enhance Manufacturing Workers' Activities." *Computers & Industrial Engineering* 163: 107804. <https://doi.org/10.1016/j.cie.2021.107804>.

Elo, Satu, Maria Kääriäinen, Outi Kanste, Tarja Pölkki, Kati Utriainen, and Helvi Kynäs. 2014. "Qualitative Content Analysis." *SAGE Open* 4 (1): 215824401452263. <https://doi.org/10.1177/2158244014522633>.

Estrada Molina, Odiel. 2022. "The Effects of WhatsApp and Telegram on Student Engagement: An Analysis from the Mixed-Methods Approach." *Education Research International* 2022: 1–14. <https://doi.org/10.1155/2022/2881404>.

Fakhriati, Fakhriati, Mu'jizah Mu'jizah, Munawar Holil, and Tedi Permadi. 2022. "Don't Leave Indonesian Manuscripts in Danger: An Analysis of Digitalization and Preservation." *Preservation, Digital Technology & Culture* 51 (1): 3–15. <https://doi.org/10.1515/pdte-2021-0017>.

Fanea-Ivanovici, Mina. 2018. "Culture as a Prerequisite for Sustainable Development. An Investigation Into the Process of Cultural Content Digitisation in Romania." *Sustainability* 10 (6): 1859. <https://doi.org/10.3390/su10061859>.

Fischer, Andreas, Emanuel Indermühle, Horst Bunke, Gabriel Viehhauser, and Michael Stolz. 2010. "Ground Truth Creation for Handwriting Recognition in Historical Documents." In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, New York, NY, USA, June 9, 3–10. <https://doi.org/10.1145/1815330.1815331>.

Flick, Uwe. 2018. *An Introduction to Qualitative Research*. London: SAGE Publications Ltd.

Hadiati, Eti, Diah Ayu Setianingrum, and Setiyo. 2024. "Information Literacy Empowerment in Islamic Education: Examining the Potential in the Indonesian Context." *Kne Social Sciences* 9 (12): 523–36. <https://doi.org/10.18502/kss.v9i12.15878>.

Haston, Elspeth, Robert Cubey, Martin Pullan, Hannah Atkins, and David Harris. 2012. "Developing Integrated Workflows for the Digitisation of Herbarium Specimens Using a Modular and Scalable Approach." *ZooKeys* 209: 93–102. <https://doi.org/10.3897/zookeys.209.3121>.

- He, Yan, Y. H., and X. R. Zhang. 2017. “‘Digital Heritage’ Theory and Innovative Practice.” *The International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences* XLII-2/W5: 335–42. <https://doi.org/10.5194/isprs-archives-xlii-2-w5-335-2017>.
- Hidalga, Nieva De La Abraham, Paul L Rosin, Xianfang Sun, Ann Bogaerts, Niko De Meeter, Sofie De Smedt, Maarten Strack van Schijndel, Paul Van Wambeke, and Quentin Groom. 2020. “Designing an Herbarium Digitisation Workflow with Built-In Image Quality Management.” *Bio-diversity Data Journal* 8: e47051. <https://doi.org/10.3897/BDJ.8.e47051>.
- Hobgood, Cherri D., Ralph J. Riviello, Nicholas Jouriles, and Glen Hamilton. 2002. “Assessment of Communication and Interpersonal Skills Competencies.” *Academic Emergency Medicine* 9 (11): 1257–69. <https://doi.org/10.1197/aemj.9.11.1257>.
- Hu, Xiao, Jeremy Ng, Shengping Xia, and Yonnie King Yiu Fu. 2017. “Evaluating Metadata Schema for Murals and Stone Cave Temples: Towards Digitizing Cultural Heritage.” In *Proceedings of the Association for Information Science and Technology* 54 (1): 491–94. <https://doi.org/10.1002/pra2.2017.14505401054>.
- Isnah, Encik Savira, Suyatno, and Heny Subandiyah. 2022. “The Effect of Digital Literacy on Language Ability in Higher Education: Experience From a Developing Country.” *Journal of Higher Education Theory and Practice* 22 (11): 215–22. <https://doi.org/10.33423/jhetp.v22i11.5425>.
- Joseph, John, and Vibha Gaba. 2020. “Organizational Structure, Information Processing, and Decision-Making: A Retrospective and Road Map for Research.” *Academy of Management Annals* 14 (1): 267–302. <https://doi.org/10.5465/annals.2017.0103>.
- Kallio, Hanna, Anna-Maija Pietilä, Martin Johnson, and Mari Kangasniemi. 2016. “Systematic Methodological Review: Developing a Framework for a Qualitative Semi-structured Interview Guide.” *Journal of Advanced Nursing* 72 (12): 2954–65. <https://doi.org/10.1111/jan.13031>.
- Knudsen, Gry H., and Marie V. Nielsen. 2019. “Exploring the Mediatization of Organizational Communication by Religious Communities in Digital Media.” *Mediekultur Journal of Media and Communication Research* 35 (66): 101–21. <https://doi.org/10.7146/mediekultur.v35i66.106274>.
- Kocak, Seyma, and Jan M. Pawlowski. 2023. “Characteristics in Digital Organizational Culture: A Literature Review.” *JKMP* 23 (2). <https://doi.org/10.62477/jkmp.v23i2.7>.
- Kumar, Srijan, and Neil Shah. 2018. *False Information on Web and Social Media: A Survey*. 1 (1). <http://arxiv.org/abs/1804.08559>.
- Kuswara, Revi. 2018. “Digital Preservation Effort of Manuscripts Collection: Case Studies of Pustakabudaya.Id as Indonesia Heritage Digital Library.” In *Lecture Notes in Computer Science*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-04257-8_20.
- Laaksonen, Salla-Maaria, and Merja Porttikivi. 2021. “Governing With Conversation Culture – Conditioning Organizational Interaction in a Digital Social Movement.” *Information Communication & Society* 25 (10): 1456–74. <https://doi.org/10.1080/1369118x.2021.1873401>.

- Lock, Irina, and Ramona Ludolph. 2019. "Organizational Propaganda on the Internet: A Systematic Review." *Public Relations Inquiry* 9 (1): 103–27. <https://doi.org/10.1177/2046147x19870844>.
- Marintan, Mega Alif, Qisthi Faradina Ilma Mahanani, and Salsabilla Nathania Arumsasi. 2024. "Digitization of javanese manuscripts in the study of archipelago manuscripts at the reksa pustaka mangkunegaran surakarta library." *International Conference on Cultures & Languages (ICCL)* 2 (1): 291–304. <https://doi.org/10.22515/iccl.v2i1.9599>.
- Masenya, Tlou M. 2021. "The Use of Metadata Systems for the Preservation of Digital Records in Cultural Heritage Institutions." *Journal of the South African Society of Archivists* 54: 55–70. <https://doi.org/10.4314/jsasa.v54i1.5>.
- Mengping, Cen. 2023. "Management and Preservation of Digital Library Resources." *The Frontiers of Society Science and Technology* 5 (17). <https://doi.org/10.25236/fsst.2023.051708>.
- Miceli, Antonio, Birgit Hagen, Maria Pia Riccardi, Francesco Sotti, and Davide Settembre-Blundo. 2021. "Thriving, Not Just Surviving in Changing Times: How Sustainability, Agility and Digitalization Intertwine with Organizational Resilience." *Sustainability* 13 (4): 2052. <https://doi.org/10.3390/su13042052>.
- Micle, Maria, Elena Tîrziman, and Angela Repanovici. 2023. "Visibility of Documentary Heritage Through Digitisation Projects in Romanian Libraries." *Plos One* 18 (1): e0280671. <https://doi.org/10.1371/journal.pone.0280671>.
- Miles, Matthew B., A. Michael Huberman, and Johnny Saldaña. 2020. *Qualitative Data Analysis: A Methods Sourcebook*. London: SAGE Publications, Inc.
- Mittal, Deepti, Rebecca Mease, Thomas Kuner, Herta Flor, Rohini Kuner, and Jamila Andoh. 2022. "Data Management Strategy for a Collaborative Research Center." *GigaScience* 12: (giad049). <https://doi.org/10.1093/gigascience/giad049>.
- Mohamad, Bahtiar, Bang Nguyen, TC Melewar, and Rossella C. Gambetti. 2018. "Antecedents and Consequences of Corporate Communication Management (CCM)." *The Bottom Line Managing Library Finances* 31 (1): 56–75. <https://doi.org/10.1108/bl-09-2017-0028>.
- Musheke, Mukelabai M., and Jackson Phiri. 2021. "The Effects of Effective Communication on Organizational Performance Based on the Systems Theory." *Open Journal of Business and Management* 09 (02): 659–71. <https://doi.org/10.4236/ojbm.2021.92034>.
- Nistane, Vaishali Chandrashekhar, and Ramprakash Brijesh Mishra. 2024. "A Brief Review of Digitization as a Modern and Technical Face of Manuscript Preservation." *Journal of Ayurveda and Integrated Medical Sciences* 9 (5): 92–6. <https://doi.org/10.21760/jaims.9.5.15>.
- Oghuvbu, Ejiroghene A., Daniel E. Gberevbie, and Samuel O. Oni. 2022. "E-Governance in Nigeria: Challenges and Prospects." *RUDN Journal of Public Administration* 9 (2): 189–99. <https://doi.org/10.22363/2312-8313-2022-9-2-189-199>.
- Olaniyi, Oluwaseun O., Jennifer C. Ugonna, Folashade G. Olaniyi, Aisha T. Arigbabu, and Chi-

nasa S. Adigwe. 2024. "Digital Collaborative Tools, Strategic Communication, and Social Capital: Unveiling the Impact of Digital Transformation on Organizational Dynamics." *Asian Journal of Research in Computer Science* 17 (5): 140–56. <https://doi.org/10.9734/ajrcos/2024/v17i5444>.

Pandey, Rahul, and Vinit Kumar. 2020. "Exploring the Impediments to Digitization and Digital Preservation of Cultural Heritage Resources: A Selective Review." *Preservation, Digital Technology & Culture* 49 (1): 26–37. <https://doi.org/10.1515/pdtc-2020-0006>.

Poltoratska, Alla, Olena Garmash, Iryna Marchenko, Олена Гарматюк, and Volodymyr Stefinin. 2024. "Analysis of the Impact of Digital Communication Tools on Organizational Culture and Communication Effectiveness Within Enterprises." *International Journal of Religion* 5 (10): 1684–92. <https://doi.org/10.61707/zaa7yf77>.

Prameswari, Clarissa, Tamara Adriani, Frans Asisi Datang, and Muhamad Prabu Wibowo. 2024. "Analisis Bibliometrik Perkembangan Penelitian Preservasi Naskah Kuno di Perpustakaan." *Jurnal Pustaka Ilmiah* 10 (1): 99. <https://doi.org/10.20961/jpi.v10i1.85220>.

Razali, Suriyati, Nor L. M. Noor, and Wan A. W. Adnan. 2009. "Structuring the Social Subsystem Components of the Community Based E-Museum Framework." In *Lecture Notes in Computer Science*, 108–16. https://doi.org/10.1007/978-3-642-02774-1_12.

Reed, Nina Petersen, and Julia Hagen. 2025. "Making People Meet: Volunteers' Contributions to Social Connection for the Well-Being of Self and Others." *International Journal of Qualitative Studies on Health and Well-Being* 20 (1): 2476215. <https://doi.org/10.1080/17482631.2025.2476215>.

Rosenthal, Meagen. 2016. "Qualitative Research Methods: Why, When, and How to Conduct Interviews and Focus Groups in Pharmacy Research." *Currents in Pharmacy Teaching and Learning* 8 (4): 509–16. <https://doi.org/10.1016/j.cptl.2016.03.021>.

Sadono, Soni, Brilindra Pandanwangi, Hening Laksani, and Danil Kasputra. 2025. "Digital Preservation of Javanese Manuscript: A Digital Asset Management Approach at Radya Pustaka Solo Museum." *KnE Social Sciences* 10 (3): 17–24. <https://doi.org/10.18502/kss.v10i3.17891>.

Sančanin, Branislav, and Aleksandra Penjišević. 2023. "The Importance of the Digitization Process for the Promotion of Cultural Heritage of the Republic of Serbia: Empirical and Statistical Findings." *International Journal of Cognitive Research in Science Engineering and Education* 11 (2): 221–29. <https://doi.org/10.23947/2334-8496-2023-11-2-221-229>.

Sidhartani, Santi. 2020. "Mapping the Variation of Visual Character Styles in the Pawukon Illustration." *IICACS : International and Interdisciplinary Conference on Arts Creation and Studies* 3: 106–11. <https://doi.org/10.33153/iicacs.v3i1.25>.

Singh, Anil K. 2012. "Digital Preservation of Cultural Heritage Resources and Manuscripts: An Indian Government Initiative." *Ifla Journal* 38 (4): 289–96. <https://doi.org/10.1177/0340035212463139>.

Sporleder, Caroline. 2010. "Natural Language Processing for Cultural Heritage Domains." *Language and Linguistics Compass* 4 (9): 750–68. <https://doi.org/10.1111/j.1749-818X.2010.00230.x>.

Suryana, Ino, Diah Chaerani, Khoirunnisa Rohadatul Aisy Muslih, Athaya Zahrani Irmansyah, Setiawan Hadi, and Anton Satria Prabuwono. 2024. "Systematic Literature Review on Optimization and Exploration of Retrieval Methods Digital Image of Ancient Manuscript as an Attempt Conservation of Cultural Heritage." *International Journal of Data and Network Science* 8 (1): 453–62. <https://doi.org/10.5267/j.ijdns.2023.9.010>.

Timofte, Radu, and Luc Van Gool. 2013. "Automatic Stave Discovery for Musical Facsimiles." In *Lecture Notes in Computer Science*. https://doi.org/10.1007/978-3-642-37447-0_39.

VOA Indonesia. 2023. "Digitalisasi Naskah Kuno: Mengabadikan Tulisan-Tulisan Dari Abad Silam." *Www.Voaindonesia.Com*. <https://www.voaindonesia.com/a/digitalisasi-naskah-kuno-mengabadikan-tulisan-tulisan-dari-abad-silam-/7111752.html>.

Wibowo, Ari. 2023. "The Influence of Government Funding on Advancing Digital Transformation within the Arts and Cultural Landscape in Indonesia." *International Journal of Education and Literature* 2 (3): 01–12. <https://doi.org/10.55606/ijel.v2i3.78>.

Wibowo, Sarwo Ferdi, Alfida, Khairul Fuad, Devi Fauziyah Ma'rifat, Fatimah Zuhrah, and Warnis. 2024. "Preservation as a Shared Responsibility: Collaboration for the Sustainable Preservation of Ambon Manuscripts." *Preservation, Digital Technology & Culture* 53 (2): 91–106. <https://doi.org/10.1515/pdte-2024-0014>.

Wikimedia Foundation. 2023. "Peluncuran Wikisource Loves Manuscripts Untuk Preservasi Digital Lebih Dari 20.000 Naskah Indonesia." <https://diff.wikimedia.org/id/2023/02/21/peluncuran-wikisource-loves-manuscripts-untuk-preservasi-digital-lebih-dari-20-000-naskah-indonesia/>.

Wikisource Bahasa Indonesia. 2024. "Wikisumber: Apakah Wikisumber Itu?" [https://id.wikisource.org/wiki/Wikisumber:Apakah Wikisumber itu%3F](https://id.wikisource.org/wiki/Wikisumber:Apakah_Wikisumber_itu%3F).

Willemse, Luc, Veljo Runnel, Hannu Saarenmaa, Ana Casino, and Karsten Gödderz. 2020. "Digitisation of Private Collections." *Research Ideas and Outcomes* 6: e57767. <https://doi.org/10.3897/rio.6.e57767>.

Zamiri, Majid, and Ali Esmaeili. 2024. "Strategies, Methods, and Supports for Developing Skills within Learning Communities: A Systematic Review of the Literature." *Administrative Sciences* 14 (9): 231. <https://doi.org/10.3390/admsci14090231>.

Zou, Changman, Sang-Yong Rhee, Lin He, Dayang Chen, and Xiaofei Yang. 2024. "Sounds of History: A Digital Twin Approach to Musical Heritage Preservation in Virtual Museums." *Electronics* 13 (12): 2388. <https://doi.org/10.3390/electronics13122388>.