Level 2: Applied | Digital Librarian

Module 09

Digital collections & discovery

AI is reshaping how patrons discover library collections, and how libraries describe them. This module covers the practical state of AI in discovery layers, digital collections, and institutional repositories.

By Yulia Brusova

20 min read · Reviewed July 2026

AI for Academic Libraries · ai-in-academic-libraries.vercel.app

WHAT YOU'LL BE ABLE TO DO

1. Identify AI features in your current discovery layer and how to configure them
2. Apply AI to improve description or accessibility in a digital collection
3. Evaluate AI transcription tools for archival or audiovisual materials
4. Draft a workflow for AI assisted finding aid enhancement
5. Understand the limitations of AI in discovery contexts
6. Evaluate newly generally available platform AI tools, including Specto and the AI Metadata Assistant for Alma, for adoption at your institution
7. Explain what AI assisted de-duplication can and cannot resolve in a shared catalog or local digital collection

> *Digital collections and discovery work sits at an unusual intersection in the current AI landscape: it is simultaneously the library function where AI has the longest operational history and the function where the most significant changes are currently underway. Machine learning has been embedded in discovery platform relevance ranking for over a decade. What is new is the ability to use generative AI to improve the underlying descriptions, transcriptions, and metadata that make discovery possible in the first place, and to do so at the scale that backlogs and resource constraints have made manually impossible for most institutions. The most useful reframe for digital collections librarians approaching AI is this: AI does not change what good discovery depends on. Good description, accurate metadata, accessible formats, and intellectually coherent organization remain the professional foundation. What AI changes is how much of that foundation a digital librarian can build with the time and staff resources actually available.*

AI in discovery layers: what has actually changed and what you can configure

Discovery platforms have incorporated machine learning for relevance ranking, query expansion, and faceted refinement for well over a decade. Primo, EBSCO Discovery Service, Summon, and WorldCat Discovery all use algorithmic relevance models that have been iteratively refined through usage data. When a patron's search returns results ranked by

relevance rather than date or call number, that ranking reflects a machine learning model, even if the platform does not describe it that way. In order to understand the current AI state of your discovery layer, the most important distinction is between features that have been present for years and genuinely new generative AI capabilities, because the professional actions appropriate to each are quite different.

The newer generation of discovery AI includes semantic search, natural language query processing, and AI generated summaries of search results. Semantic search interprets the meaning of a query rather than matching its exact terms, which means a patron searching for "how stress affects memory" retrieves conceptually relevant results even when sources use different vocabulary such as "cortisol and hippocampal function" rather than the patron's own phrasing. Harvard University Library's Collections Explorer platform, launched in 2024, provides a concrete institutional example: it allows researchers to explore Harvard's collections through natural language queries, combining semantic search with generative AI to surface connections across disparate collection areas that keyword search would not reveal. Such a shift in patron experience, from search term construction to conversational inquiry, is not merely an interface change. It reflects a fundamental change in how the discovery layer interprets patron intent.

OCLC added AI assisted classification and subject heading suggestions to its WorldShare Record Manager and Connexion cataloging applications in December 2025, drawing on WorldCat's hundreds of millions of bibliographic records to suggest Dewey Decimal Classification numbers, Library of Congress Classification numbers, and Library of Congress Subject Headings at the point of cataloging. Such a feature illustrates the direction all major platform vendors are moving: AI embedded directly in the professional workflow rather than offered as a separate tool. Furthermore, Ex Libris launched Specto, an AI powered platform for digital collections management, which incorporates AI metadata creation, bulk editing, and discovery enrichment into a single integrated workflow.

Specto is now generally available, alongside a companion tool worth the same attention: the AI Metadata Assistant for Alma, which works inside Alma's cataloging interface to draft descriptive metadata, suggest subject headings, and flag likely duplicate records during import and ingest. Both tools sit under Ex Libris's broader Academic AI platform, the umbrella under which the company is consolidating the AI capabilities it is building across Alma, Primo, and Rosetta. For digital collections librarians at institutions running Ex Libris products, general availability changes the relevant professional question. Specto and the AI Metadata Assistant for Alma are no longer features to watch for in a future release; they are features to request access to, configure for a pilot collection, and evaluate against the audit and pilot framework this module returns to later. In order to make that evaluation meaningful, the digital librarian

needs to know what the tools actually do with a specific institution's records, which is a question only a pilot on local data can answer.

The professional limitation that matters most in this landscape is the distinction between vendor controlled AI features and locally configurable ones. Many discovery platform AI capabilities are determined by the vendor and applied uniformly across subscribing institutions; a digital librarian cannot adjust the semantic model or the relevance weights directly. What is configurable at the local level varies by platform and contract: relevance tuning profiles, query expansion settings, local boosting rules, and which AI features are exposed to patrons. In order to know what is actually available at your institution, the most direct approach is to review your platform's current release notes, contact your vendor representative with specific questions about AI features enabled in your subscription tier, and request a demonstration of any semantic search or AI summary capabilities that are available but not yet turned on. Such advocacy for specific features, rather than passive acceptance of the platform's defaults, is a genuine professional function that improves patron experience in ways the vendor cannot accomplish without institutional input.

Semantic search and natural language discovery: the patron experience is shifting

The practical difference between keyword search and semantic search is visible in patron behavior, even when patrons cannot articulate the difference. A patron who types a natural language question into a semantically capable discovery layer and receives relevant results is experiencing something fundamentally different from Boolean search, even if the interface looks similar. Such a patron is not required to know controlled vocabulary, translate research questions into search terms, or apply Boolean operators, and for patrons who have spent their lives seeking information through conversational search engines, the absence of that translation requirement is significant.

Yale Library has developed a prototype for a Digital Collections AI application that applies large language models directly to digitized texts, performing summarization, translation, named entity extraction, and question answering on items from Yale's digital collections. For example, a researcher working with digitized historical correspondence could ask the system a question about the correspondence's content and receive a synthesized answer drawn from the transcribed text, rather than needing to read every letter to locate a specific reference. Such a capability transforms the researcher's relationship to a large, underdescribed collection: instead of navigating finding aids and reading item by item, the researcher can interact with the collection's content directly. Yale is actively seeking

instructor partnerships to test the application, which means that libraries investigating similar capabilities have a peer institution example to examine and a potential collaboration to pursue.

Knowledge graphs represent a related but distinct approach to discovery enhancement. Tools like Yewno connect concepts across resources using semantic relationships rather than keyword matching, allowing a patron to follow a concept through related ideas, disciplines, and sources in ways that traditional subject heading navigation does not support. Primo's "Topics" feature uses a similar underlying logic, surfacing conceptual relationships between a search result and connected ideas based on how concepts are linked across the literature. Such relational discovery is particularly useful for interdisciplinary research, where a concept in one discipline has meaningful connections to literature in adjacent fields that might not share controlled vocabulary.

The professional responsibility that semantic search creates is instruction related rather than technical. Patrons who use natural language discovery tools successfully may conclude that AI search is more comprehensive than it is, reasoning that if the tool did not surface a source, the source does not exist or is not relevant. In order to prevent such a misconception from affecting patron research quality, instruction should address the difference between semantic discovery and comprehensive bibliographic retrieval: AI can surface likely relevant results, but it cannot guarantee completeness in the way a systematic database search with explicit inclusion criteria can. Additionally, semantic search results require the same evaluation that keyword search results require; relevance ranking does not confer authority, and a result appearing at the top of an AI ranked list is not thereby more credible than one appearing lower. Such instruction points should be integrated into any session where patrons are introduced to a semantically capable discovery platform.

Intellectual freedom in AI discovery: transparency, opt-out, and non-personalized access

The shift from keyword search to AI-mediated discovery is not only a usability improvement; it is a change in who, or what, decides which materials a patron sees and in what order, and that makes it an intellectual freedom question as much as a technical one. ALA's AI guidance names Intellectual Freedom as a core value precisely because AI systems threaten it when they, in the guidance's words, "obscure how information is ranked, extract public or library-created data without clear public benefit, or optimize engagement at the expense of attention." A discovery layer whose relevance model is a vendor-controlled black box is doing exactly the first of these: it is ranking the collection by criteria neither the librarian nor the

patron can inspect, and a patron who does not know why certain results rose to the top cannot reason about what the ranking might be burying. In order to take this value seriously, a digital collections librarian has to treat the discovery layer's ranking not as neutral plumbing but as an editorial act that requires accountability.

Three concrete obligations follow from the guidance, and each is actionable at the level of discovery configuration and instruction. The first is transparency: the guidance calls for "clear documentation explaining automated system function, purpose, and accountability," which means a library should be able to tell patrons, in plain language, that discovery results are algorithmically ranked and personalized, and should press its vendor for enough explanation of the ranking model to make that disclosure honest. The second is the preservation of predictable, non-personalized paths. The guidance asks libraries to "preserve predictable, transparent, non-personalized access paths alongside AI-enhanced discovery," which in practice means keeping a straightforward, non-AI way to browse and search the collection available rather than replacing it entirely with a personalized, engagement-optimized interface. For example, a patron who wants to see everything the collection holds on a topic, in a stable and reproducible order, is exercising a legitimate research need that a purely semantic, personalized discovery layer can actively frustrate, and preserving a non-personalized path protects that need. The third is opt-out: the guidance holds that "users must have opt-out capability when possible" and that vendors should "offer equivalent non-AI services or libraries document limitations," which turns the availability of a non-AI route into a procurement question a library should raise before adoption, not a feature it hopes the vendor happens to provide.

The most distinctly intellectual-freedom obligation, and the one easiest to overlook, is the audit for suppressed visibility. The guidance asks for "regular audits of discovery and recommendation features for patterns reducing visibility," which is a different exercise from evaluating relevance quality. A relevance audit asks whether the top results are good; a visibility audit asks whether certain materials, certain viewpoints, or certain communities are being systematically pushed so far down the ranking that they are effectively invisible, regardless of how good the top results look. For example, a discovery layer that consistently ranks recent, high-usage, English-language materials at the top may, without anyone intending it, bury older holdings, minority-language materials, and low-circulation works documenting marginalized communities, and only a deliberate audit that looks for what is being hidden rather than what is being surfaced will catch it. Such an audit connects directly to the bias-auditing practice that Module 05 describes, and it shares the same premise: a pattern that disadvantages certain materials is invisible in any single search and detectable only across many. What the guidance does not do is convert these obligations into a

checklist a system can satisfy on its own. Deciding that a discovery configuration honors intellectual freedom, that its ranking is accountable, its non-AI paths genuine, and its visibility patterns fair, remains a professional judgment that belongs to the librarians responsible for how their community reaches the collection.

Finding aids and AI assisted description: making archival collections accessible

The description gap in archival collections is one of the most persistent practical challenges in digital library work. Collections that were digitized under grant funding, often with tight timelines and minimal staffing, frequently have finding aids that contain collection level scope notes and series descriptions but no item level description at all. Such collections are technically accessible but practically invisible: a patron who cannot find an item through description cannot know it exists, regardless of how thoroughly it has been digitized. AI description assistance addresses this gap at a scale that manual remediation, with existing staff and resource levels, cannot match.

The workflow for AI assisted finding aid enhancement proceeds from existing content rather than generation from scratch. For a collection with a collection level scope note, series descriptions, and digitized item images or transcriptions, AI can draft plain language summaries of each series, generate patron accessible overviews of complex intellectual arrangements, and produce item level descriptions from provided transcription content. For example, an archivist working with a digitized correspondence collection can provide AI with a letter's transcription and prompt it to write a three to five sentence catalog description identifying the sender, recipient, date, primary subject, and any significant details a researcher would find relevant. Such a description is a draft requiring archivist review for factual accuracy, appropriate terminology, and adherence to archival description standards, but it is a draft produced in seconds rather than minutes, which represents a substantial per item efficiency gain when applied across a collection of several hundred items.

The particular strength of AI in finding aid work is translating archival language into patron accessible prose. Finding aids written in traditional archival description conventions, with terminology like "accruals," "extant," "bulk dates," and "provenance," are opaque to most patrons who have not received archival research training. For example, providing AI with an existing scope and content note and asking it to rewrite the note for an undergraduate student who has never visited an archive before produces a genuinely more accessible document with the same information presented in a different register. Such plain language

rewriting preserves the intellectual content of the description while substantially expanding the patron population who can interpret it.

Heightened professional care is required for two categories of archival material. First, collections containing records about living individuals, including recent institutional records, community archives, and personal papers with donor imposed restrictions, require careful review to ensure AI generated descriptions do not disclose information beyond what the access policy permits. AI does not inherently know which details in a document are sensitive; the archivist's professional judgment about appropriate description remains the only reliable filter. Second, collections documenting Indigenous communities and cultural materials require description that reflects the originating community's preferred terminology and the collection's access conditions, which may be community governed rather than institutionally determined. AI trained on mainstream archival description conventions will not automatically apply Indigenous centered description principles; such principles must be provided explicitly in the prompt and verified in the output. Additionally, any finding aid content generated with AI assistance should be disclosed in the collection's processing note, both as a matter of professional transparency and as documentation useful for future revision.

Accessibility: AI transcription, image description, and removing barriers at scale

The accessibility debt carried by most digital collections programs is substantial and largely invisible until a patron with a disability encounters it. Audio and video materials without transcripts are inaccessible to deaf and hard of hearing users. Images without alt text are inaccessible to blind and low vision users who rely on screen readers. Digitized text with poor OCR quality is inaccessible to users who depend on text to speech tools. Section 508 of the Rehabilitation Act requires accessible electronic content for federal agencies and many federally funded institutions, and most digital collections programs, despite best intentions, have not been able to generate the accessible formats that compliance requires because the manual labor involved exceeds available staff capacity.

AI transcription has reached a level of reliability that makes it a practical tool for production workflows rather than an experimental one. OpenAI's Whisper model, which is open source and free to run locally, produces transcripts from audio files with accuracy levels sufficient for most library applications after human review. For oral history collections (where interview recordings may run several hours each and manual transcription would require paid professional services or substantial volunteer labor), AI transcription converts an economically prohibitive task into a review and correction workflow. For example, a digital

librarian managing a collection of fifty oral history interviews, each averaging ninety minutes, faces approximately seventy five hours of manual transcription to provide accessible transcripts. Running those recordings through Whisper and reviewing the resulting transcripts takes considerably less time, particularly for recordings with good audio quality and standard English speech patterns. Such a reduction in labor barrier is what allows accessibility compliance to become practically achievable rather than permanently deferred.

Accuracy varies predictably with audio quality, speaker accent, technical vocabulary, and background noise. Recordings with poor audio, multiple simultaneous speakers, strong regional accents, or specialized disciplinary terminology will produce transcripts requiring more substantial review and correction. Such variation is a reason to pilot transcription on a representative sample of recordings before committing to a production workflow; it allows the digital librarian to calibrate the review time required for a specific collection's characteristics before planning a full scale project. Furthermore, for collections documenting languages other than English, both Whisper and commercial tools like Otter.ai and Rev AI show variable performance that requires evaluation against the specific language before the workflow is adopted.

AI alt text generation for images in digital collections addresses a second major accessibility gap. For image collections such as photographs, maps, illustrations, and digitized artworks, AI can generate descriptive alt text that screen reader users receive in place of the visual content. The practical workflow involves providing the image to a vision capable AI model and requesting a description appropriate for the image type and patron use context. For example: "Write alt text for this photograph from a 1950s university yearbook. Describe the image content precisely in one to two sentences. Do not interpret or editorialize; describe what is visible." Such a prompt produces draft alt text that the digital librarian reviews for accuracy and appropriate detail level. The review step is essential: AI image description can misidentify individuals, misread text within images, or provide descriptions that are accurate but not optimally useful for the research context. Such errors are reliably catchable through human review, and catching them before publication is both professionally and institutionally important.

Institutional repositories: addressing metadata debt with AI

Institutional repositories accumulate metadata inconsistency the way physical collections accumulate dust: gradually, invisibly, and in ways that become expensive to address only once the accumulation is large enough to affect discoverability in measurable ways. Author name variants arise from inconsistent formatting across departments and submission

periods. Subject terms reflect practices from multiple eras and multiple catalogers. Abstracts are missing from older deposits that predate the repository's current deposit workflow. Institutional affiliation fields contain every variation of the institution's name that has been used over its history. Such inconsistencies reduce discoverability both within the local IR interface and in aggregated indexes like BASE, CORE, and OpenDOAR, which rely on metadata quality for reliable syndication.

AI batch processing addresses IR metadata inconsistency at the scale that manual remediation cannot. The approach follows the same information in logic that makes AI batch processing reliable in cataloging contexts: AI given a clear normalization task and a representative set of records applies rules consistently across the full set, which the digital librarian reviews for quality before committing to the production data. For example, a digital librarian managing an IR with five thousand records carrying inconsistent author name formatting can export a CSV of affected name fields, provide AI with the target format and three to five examples of the transformation, and receive a corrected dataset for review. Such a task, done manually, might require days of careful review and editing. Done with AI assistance and post processing review, it can typically be completed in hours, with the important caveat that the review step must be taken seriously, not treated as a formality.

Subject term enrichment for poorly described IR deposits represents a second high value application. Records deposited without subject terms, or with only local, nonstandard terms, are difficult to discover through standard subject facets and fail to syndicate effectively to aggregated indexes that rely on controlled vocabulary. Providing AI with a deposit's title, abstract, and any available keywords and asking it to suggest appropriate subject terms from a specified vocabulary produces candidate terms that the digital librarian evaluates against the actual controlled vocabulary and the item's content. For example: "Suggest five to seven subject terms for the following dissertation abstract, using Library of Congress Subject Headings. Note any terms you are uncertain about." Such a prompt produces a working set of suggestions that substantially reduces the per record effort of subject enrichment for a large backlog.

Ex Libris has implemented AI metadata enrichment directly in Alma for ProQuest EBook Central records, automatically adding language, summary, and subject heading fields in alignment with Library of Congress standards for records in the Alma Community Zone. Such a vendor implemented example illustrates the direction the field is moving: AI metadata enrichment becoming a standard feature of IR and library management platforms rather than a locally implemented workflow. For digital librarians evaluating platform upgrades or new IR systems, AI metadata capabilities are now a relevant evaluation criterion, and a question worth asking any vendor whose platform manages institutional repository deposits.

AI assisted de-duplication: what the OCLC WorldCat test demonstrates

Duplicate records are one of the quieter forms of metadata debt, and one of the most consequential for discovery. In a shared cataloging environment like WorldCat, duplicate records for the same item split holdings information across multiple records, which means a library's holding may attach to a record other patrons and other libraries never see. In a local digital collection or institutional repository, duplicates arise from re-deposits, harvesting the same item from multiple source systems, or records created independently by different staff members before a backlog was addressed. Such duplicates inflate result counts, confuse patrons who encounter the same item twice in a results list, and complicate any downstream use of the collection's metadata, including syndication to aggregated indexes.

On February 11, 2025, OCLC ran a test of AI assisted de-duplication against approximately 500,000 record pairs of print English-language books in WorldCat, evaluating how well an AI model could identify which pairs represented true duplicates of the same edition and which represented distinct but closely related records, such as different printings, different editions, or multi-volume sets with similar titles. A test at that scale matters to digital collections librarians for the shape of the problem it illustrates, not only its size. Candidate duplicate pairs in a collection this large separate naturally into two categories: pairs where two records unmistakably describe the same edition of the same book, and harder pairs where title, author, and publication information are similar but the records describe a different edition, a different printing, or a multi-volume set that should remain distinct. The second category is where review matters most, because a merge that collapses two genuinely different editions into one record destroys information a researcher might need, such as which edition contains a particular revision or which printing a specific provenance note refers to.

The same logic transfers directly to local digital collections and institutional repository work, at a smaller scale but with the same stakes. For example, a digital librarian who suspects an institutional repository contains duplicate deposits of the same thesis, submitted once at deposit and again during a later metadata cleanup project, can use AI to compare records across the repository and flag pairs that share enough metadata in common to warrant review, producing in an afternoon a candidate list that manual comparison of several thousand records would take considerably longer to produce. What AI assisted de-duplication does not do, in either the WorldCat scale test or a local repository, is make the merge decision itself. Determining whether two records describe the same intellectual object, or two objects that merely resemble each other in their metadata, remains a judgment that

depends on examining the actual items, not only their descriptions, and that judgment belongs to the cataloger or digital librarian reviewing the AI's candidate list, every time.

Building a sustainable AI enhanced digital collections workflow

The most common mistake in AI adoption for digital collections work is selecting the tool before defining the problem. A digital librarian who reads about AI image description and immediately begins generating alt text for the most prominent collection in her repository may find that the workflow produces inconsistent results, requires more review time than anticipated, and is difficult to hand off to colleagues because the process was developed informally. In order to build a workflow that scales and sustains, the sequence should move from audit to pilot to documented process, with the documentation step treated as essential rather than optional.

The audit phase begins with a discerning assessment of where the most significant description, accessibility, and discoverability gaps exist in the current digital collections program. For most institutions, this means answering three questions: Which collections have the most serious description deficits? Which materials lack accessibility equivalents that are required or most needed? Where is metadata inconsistency most affecting patron discoverability? Such an audit does not require a comprehensive collections inventory; it requires enough knowledge of the collection landscape to identify the highest priority targets for AI assistance. The digital librarian who has managed a repository for several years typically already knows where the most significant problems are; the audit formalizes that knowledge into a prioritized list that can guide a phased implementation.

The pilot phase applies AI assistance to a small, bounded scope such as one collection, one record type, or one batch of fifty items, before committing to a full scale workflow. Such a pilot serves two purposes: it reveals the practical quality level of AI output for the specific collection characteristics and prompt design being used, and it generates realistic estimates of review time that can inform project planning. For example, a digital librarian piloting AI transcription for an oral history collection might run ten recordings through Whisper and time the review and correction process for each, then use that data to estimate the full collection's transcription timeline. Such empirical data is more reliable than vendor estimates or peer institution benchmarks, because it reflects the specific characteristics of the actual recordings (audio quality, speaker patterns, technical vocabulary) rather than generic performance claims.

The documentation step is where sustainable workflow development most often breaks down. A workflow that works for one person and has never been written down is not an

institutional workflow; it is a personal practice that disappears when the person changes roles or institutions. In order to build a workflow that persists and can be handed off, the documentation should capture the prompt or prompt template used, the quality review criteria applied, the record keeping conventions for flagging AI assisted records, and the conditions under which the workflow requires escalation to additional professional review. Such documentation is the difference between AI adoption that builds institutional capacity and AI adoption that builds individual efficiency, and for digital collections programs managing long term collections, institutional capacity is the appropriate goal. Furthermore, sharing workflow documentation with peer institutions through professional networks such as the Digital Library Federation, the Society of American Archivists, and OCLC communities, contributes to the shared professional knowledge base that makes adoption easier for institutions earlier in the process.

PRACTITIONER NOTE

The accessibility application is what demonstrates most clearly that AI belongs in digital collections workflows, not as a productivity enhancement but as a genuine service expansion. An oral history collection with over sixty recordings and no transcripts represents exactly the kind of accessibility debt that has accumulated for years because manual transcription is financially out of reach. Running the first ten interviews through Whisper and reviewing the output takes an afternoon. The transcripts are not perfect: there will be proper noun errors, some regional vocabulary the model does not recognize, and a few passages that need significant correction. But they are usable after review, and they are better than no transcript. What this illustrates is not merely efficiency, though the efficiency is real. It is that the barrier between 'this collection exists' and 'this collection is accessible' has dropped low enough to actually cross. That is a different category of value than saving time on a task that was already being done.

KEY TAKEAWAYS

- Discovery platforms have incorporated machine learning for relevance ranking for over a decade; genuinely new capabilities include semantic search, natural language query processing, and AI generated summaries, most of which are vendor controlled rather than locally configurable.
- Semantic search and natural language discovery change the patron experience by interpreting query meaning rather than matching terms; instruction should address

both the expanded access this provides and the completeness limitations it does not resolve.

- AI-mediated discovery is an intellectual freedom question: ALA's guidance asks libraries to disclose that results are algorithmically ranked, preserve predictable non-personalized access paths and an opt-out alongside AI discovery, and regularly audit discovery and recommendation features for patterns that reduce the visibility of certain materials.
- AI assisted finding aid description is most reliable when working from existing content (scope notes, transcriptions, partial records) rather than generating description without provided material; sensitive collections require heightened archivist review before any AI assisted description is published.
- AI transcription using tools like Whisper makes accessibility compliance achievable for oral history and AV collections where manual transcription is economically prohibitive; accuracy varies with audio quality and should always be reviewed before publication.
- Institutional repository metadata debt (inconsistent author names, missing subject terms, variable description quality accumulated across deposit years) is addressable through AI batch processing applied with explicit normalization rules, followed by quality review before production commitment.
- Specto and the AI Metadata Assistant for Alma, both part of Ex Libris's broader Academic AI platform, are now generally available; for institutions running Ex Libris products, the relevant step is no longer awareness but a local pilot evaluating these tools against real collection data.
- OCLC's February 2025 test of AI assisted de-duplication on roughly 500,000 WorldCat record pairs of print English-language books showed AI can reliably flag candidate duplicates at scale, but the harder cases, distinct editions, printings, and multi-volume sets, still require a cataloger's review before any merge, a pattern that applies equally to de-duplication in local digital collections and institutional repositories.
- A sustainable AI enhanced digital collections workflow requires three phases in sequence: an audit identifying the highest priority gaps, a pilot on a bounded scope that generates realistic quality and time estimates, and documented processes that build institutional capacity rather than individual efficiency.

References

APA 7TH EDITION

1. American Library Association. (2026). *Guidance on the use of artificial intelligence in libraries.* https://www.ala.org/tools/standards-and-guidelines/guidance-use-artificial-intelligence-libraries
2. Association of College and Research Libraries. (2025, October). *AI competencies for academic library workers.* American Library Association. https://www.ala.org/acrl/standards/ai
3. Ex Libris. (2025). *Alma Specto: AI-powered digital collections management.* Clarivate. https://exlibrisgroup.com/products/alma-library-services-platform/
4. Harvard Library. (2024). *Harvard Library Collections Explorer.* Harvard University. https://library.harvard.edu

   Confirm direct URL for the Collections Explorer.
5. Yale University Library. (2024). *Digital collections AI application* [Prototype]. Yale University. https://library.yale.edu

   Confirm direct URL for the Yale digital collections AI application.

ACRL AI Competencies covered

Use & Application

Sub-competencies: 4.1, 4.4, 3.2 · ACRL AI Competencies (2025)