



Large-Scale Digitization in the AI Era

Issues and Model Terms for Libraries and Their Partners

September 9, 2026: Public Discussion Draft

Center for Research Libraries www.crl.edu

Large-Scale Digitization in the AI Era

Issues and Model Terms for Libraries and Their Partners

This briefing document was developed in response to discussion among CRL member library directors about the terms that libraries should seek when collections are proposed for large-scale digitization, particularly where the resulting digital content and metadata may support commercial AI or other proprietary services. This public version is being released to support all libraries in navigating these issues, which CRL members see as a matter of broad professional concern. CRL's primary role at this time is to help libraries make informed decisions by gathering experience and understanding member questions in order to develop baseline terms for partnership agreements. CRL intends for this to lead to practical action, but the particular courses of actions to be taken—from negotiations over any specific partnership, pooled requests for proposals or similar means of brokering commercial use of library content, or developing a shared corpus under specific model terms—require further member discussion and legal review.



Contents

Summary	1
Background	2
Key Decision Points	4
Questions for Institutional Consideration	5
Working FAQ	6
Is CRL advising for or against any specific partnerships?	6
Why not simply say no?	6
Is this about selling collections or finding a market price?	6
Does receiving and “owning” a digital copy solve the rights question?	6
Does metadata have independent value?	6
How does open access fit with concern about AI harvesting?	7
What is CRL’s role at this stage?	7
Model Digitization Terms	8
About the Center for Research Libraries	11



Summary

CRL and many of its member libraries have been approached about large-scale digitization projects by AI developers. Large-scale digitization can enhance discovery, access, accessibility, research, and preservation. It can also generate substantial commercial and strategic value, particularly when library collections or metadata are used to develop AI systems.

CRL supports genuine partnerships that advance the mission of libraries, improve AI for research, and help libraries understand how their distinctive resources can contribute to these technologies. To this end, CRL is developing model terms that account for the range of member interests at play and will assess any partnership proposed to CRL against them.

This document explains CRL's approach and has been released for public use to help all libraries evaluate pertinent factors in digitization agreements, obtain partnership terms aligned with their individual mission and goals, and act in accord with shared interests across our profession. This is a developing issue area and CRL will update this document based on input from its members and its own experience developed in navigating these issues.

This document is intended to support, not impede, effective large-scale digitization. Libraries have existing obligations created by contracts and grants and good reasons to pursue programs of digitization in the course of user services. Libraries have promoted technological development and benefited from it. They have also had to accommodate and adapt to both the direct costs and downstream consequences of new technologies.

The development of AI makes this a pivotal moment for libraries to consider the terms of digitization deliberately: materials digitized and made available online for any purpose may reasonably be expected to become inputs to AI development, whether or not that was the original intent of the project. Careful evaluation of the goals of partnership and the terms that protect those goals is crucial to ensuring that AI development yields the best obtainable value for libraries.

Background

Large-scale digitization in the AI era requires a broader assessment than libraries have previously applied. Access and stewardship outcomes still matter but libraries also need to assess the types of value their materials may create and the effects of AI on the communities they serve. Although difficult to control or predict, libraries should consider whether the likely downstream uses for digitized content advance the library's mission or create negative externalities. A fair partnership should return proportionate value, account for both direct and long-term stewardship costs, and provide transparency about how content will be used, in order to support assessment of the project's wider effects on the information ecosystem.

The interests of libraries and their stakeholders have not been clearly mapped onto the economics of AI. National media and the financial press have reported commitments for datacenters approaching \$1 trillion (USD) across recent years, and frontier development labs have reported tens of billions (USD) in annual revenue. This economic activity is enmeshed with training AI on digital content, yet factors that matter to the value of library collections—such as rarity, scholarly merit, or comprehensiveness—do not have a clearly understood relationship to commercial value for AI development. Also unknown is the degree to which exclusive control over any body of content, for AI development or use with AI systems, will yield a commercial advantage.

While acknowledging that questions about how library collection assessment factors relate to AI systems' performance or market value remain open, CRL's position is that competition among AI services should be driven by technical merit, not by restrictions on access to information. It is crucial to sustain libraries' ability to make content in their care available for non-commercial, scholarly and academic, research and teaching applications. For CRL, this requires care so that agreements do not prohibit it from filling member requests for use of the shared collection in conformity with CRL's mission and exempt purposes. More broadly, the copyright and fair use landscape for AI development remains unsettled, so even robust indemnification provisions may not fully offset the legal and reputational risks libraries face.

Libraries need to take an active role and critical stance at a moment when AI is recasting the value and sociopolitical importance of their role in maintaining the record. As AI increases the volume of claims requiring validation, the need for independent stewards of authoritative sources becomes more important. Probabilistic systems can produce accurate information

and relevant citations, but they cannot themselves preserve the underlying record or guarantee its authority and provenance.

Libraries hold a distinctive resource in their analog collections and should be deliberate about decisions to transform those materials into digital commodities. CRL and its members are already seeing patrons turn to libraries to verify AI-generated claims and pursue new research directions prompted by these tools. Libraries are hard at work developing AI literacy standards and educational programs, an extension of the critical role they play in representing and advocating for the needs of their researcher communities. This suggests AI may increase demand for library expertise and services, which carry a real cost, even as library collections are used to build commercial technologies.

Libraries face both ethical and practical challenges at this moment. Our profession has an abiding commitment to support research without judging its purposes, while protecting patron privacy. Libraries generally take pains to create the minimum possible record of who is using their content and avoid restrictions based on the purposes of its use. Patron privacy and non-discrimination in access do not imply an obligation to make the library itself indifferent to enclosure, market power, or uncompensated commercial appropriation.

Developing model terms protects the field's overall negotiating position. Individual agreements with weak terms can set a market norm and can reduce the leverage of other institutions. Shared principles and informed discussion therefore matter even when libraries ultimately make their own decisions. In counterpoint, although CRL's assessment is that a pause on new partnerships is the best course of action, it is reasonable to assume the value of library collections for AI development is time-bound, so delays should be minimized.

These recommendations are not opposition to commercial collaboration or AI development. Libraries have a long history of working with publishers and other commercial partners to sustain education and research, and AI should be brought into that tradition on fair and mission-aligned terms. To support that work, CRL has prepared a model terms sheet and a set of frequently anticipated questions for libraries to consider.

Key Decision Points

The following points have informed CRL's position and are offered as a guide for conversations within the library and with stakeholders on campus or in the community about the role of the library in relationship to AI development.

- Mission and public interest: What access, preservation, scholarly, or broader public purpose will the project advance?
- Library rights: Does the library retain authority over scope and broad lawful uses of the resulting files and metadata that it obtains from the partnership?
- Reputational factors: Does the library have control over how its affiliation with the project will be described?
- Clarity of partner use: Are commercial, AI, derivative, affiliate, and post-project uses stated affirmatively rather than left implicit? Is the license exclusive or non-exclusive?
- Reciprocal value: Do benefits to the library, its users, and the public reasonably correspond to the burden and value contributed?
- Long-term stewardship: Do technical, physical-handling, metadata, access, and legal terms preserve future options rather than create lock-in or risk to collections?

Questions for Institutional Consideration

The role of libraries in relation to AI is a complex and developing subject. These questions are intended to prompt discussion and guide conversations within libraries and with their stakeholders.

1. Under what conditions should research libraries participate in large-scale or AI-adjacent digitization?
2. Are there different factors that bear on commercial partnerships and non-commercial or academic research projects?
3. Which rights should always remain with the contributing library, regardless of who pays for digitization?
4. What forms of reciprocal value best sustain access, preservation, description, and long-term stewardship?
5. Which partner uses require explicit authorization, reporting, time limits, or additional consideration? How will new use cases be evaluated and ambiguities resolved?
6. How should libraries reconcile open access for individual users with large-scale machine extraction and commercial reuse? What systems are warranted to help protect open access from commercial exploitation?
7. Where would shared standards, coordinated discussion, or collective action be more effective than institution-by-institution negotiation?
8. What prior agreements, operational experience, or institutional constraints should inform CRL's model and next steps?

Working FAQ

Is CRL advising for or against any specific partnerships?

No. This brief is partner-agnostic. The immediate purpose is to educate, gather member experience, and develop shared expectations around model terms for individual agreements or a more coordinated structure. CRL's recommendations in this document reflect our overall judgment about negotiation strategy in a complicated and consequential area.

Why not simply say no?

Libraries have a mission interest in improving the information environment their communities actually use, and a rational economic interest in funding that mission. The issue is whether participation obtains the best value for the library while avoiding mission conflict or unacceptable risks.

Is this about selling collections or finding a market price?

No. The salient matter at stake is sustaining access and stewardship, and ensuring the library can act effectively on behalf of its community. Cash may be appropriate, but value may also take the form of in-kind benefits such as digitization, description, preservation, accessibility, infrastructure, or other durable public and member benefit.

Does receiving and “owning” a digital copy solve the rights question?

No. First and foremost, no agreement around derivatives of these projects can circumvent copyright or other relevant intellectual property law. Ownership of a designated group of digital files can coexist with contractual limits on display, sharing, preservation, computational use, or access. Permitted uses matter more than the assertion of ownership.

Does metadata have independent value?

Yes. Cataloging, description, identifiers, and other metadata reflect substantial professional and institutional investment and may themselves support AI and commercial products. Libraries must be especially conscious of administrative and behavioral metadata, which may have significant value for AI development because it reveals patterns of use, interaction, and decision-making. Use of all types of metadata should be considered as part of the value the library offers, and thus expressly authorized and appropriately limited.

How does open access fit with concern about AI harvesting?

This remains a serious and unresolved tension. Open access for individual users does not necessarily answer questions about bulk machine extraction and the concomitant operating costs, implications of commercial reuse, or exclusivity of downstream products. Agreements and local system designs may need different expectations for different forms of use.

What is CRL's role at this stage?

The first role is convening and education: develop baseline terms, gather experience and understand member questions, and help libraries make the best decisions possible. CRL intends for this to lead to practical action, but the particular courses of actions to be taken—from negotiations over any specific partnership, pooled requests for proposals or similar means of brokering commercial use of library content, or developing a shared corpus under specific model terms—require further member discussion and legal review.

Model Digitization Terms

The following standards describe positions CRL would generally seek in a large-scale or AI-adjacent digitization agreement. They are a negotiating framework, not a complete contract, and do not constitute legal advice. As CRL gathers input from its members and completes further evaluation, it will update its posture and baseline requirements for partnership and key areas for negotiation.

Area	Model standard	Posture
Purpose, scope, and governance	State the public-interest purpose, intended users, selection rationale, collection scope, exclusions, project plan, oversight, reporting, and change-control process.	[Baseline: The library retains final authority over what is to be digitized.]
Library rights and permitted uses	The library receives complete preservation-quality files and may preserve, migrate, host, reproduce, analyze, index, describe, make accessible, lend, transmit, and otherwise use them for lawful mission purposes.	[Baseline: Members, researchers, contractors, and service provider use may be authorized under library policies.]
Partner and AI uses	State affirmatively whether and how the partner may use images, OCR, metadata, embeddings, datasets, annotations, or derivatives. AI training, evaluation, retrieval, model improvement, synthetic data, and commercial data services require express, purpose-specific authorization.	[Baseline: Agreement is project-specific rather than a blanket authorization] N.B. This area requires affirmative commitments, so that silence is not taken as permission.
Affiliates, transfer, and survival	Limit rights to the contracting entity and disclosed affiliates or subcontractors bound by equivalent terms. Prohibit onward sale or sublicensing of the library-derived data except as approved. Define what ceases, is deleted, or may survive after termination or change of control.	[Baseline: Limit rights to the contracting entity and disclosed affiliates or subcontractors bound by equivalent terms.]; duration variable.

Area	Model standard	Posture
Public access and member benefit	Provide no-cost access to public-domain materials without avoidable barriers. Define benefits for affiliated institutions, consortia members, and individuals, such as content access, metadata, tools, computational access, accessibility outputs, training, or credits.	[Baseline: individual non-commercial researchers should be free to use public domain materials]; benefits sought may vary.
Metadata and knowledge infrastructure	License only the minimum amount of metadata needed to identify the works for the approved project. More comprehensive descriptive metadata have additional value. Avoid warranties concerning third-party records. Require return of enriched metadata, OCR, layout, identifiers, provenance, and quality data in documented, reusable, interoperable formats.	[Baseline: library provides a minimum amount of metadata needed for executing the project]; benefits sought may vary; specifications are essential and may be included in operational plans so long as those cannot be unilaterally changed. N.B. Providing rich metadata for identifying collections of interest may be warranted; the baseline aims to protect the value of metadata in the subsequent project.
Digital technical standards	Define digitization specifications appropriate to the source and intended use, including resolution, color, compression, OCR, file naming, packaging, checksums, manifests, accessibility outputs, and acceptance thresholds.	[Baseline: Require preservation masters and correction or reprocessing of defective work at the partner's expense.]; specifications may vary and may be included in operational plans so long as those cannot be unilaterally changed.
Physical handling and logistics	Separately define packing, transport, custody, environmental conditions, tracking, conservation review, insurance, loss and damage responsibility, incident reporting, and authority to halt work.	[Baseline: Institutionally specific]; logistical requirements may vary and need to be determined by institutional standards. For CRL and other print archives, the physical integrity of collections cannot be compromised for digitization and CRL urges all libraries to evaluate scarcity and work with print archiving programs to ensure copies are retained in reliable repositories.

Area	Model standard	Posture
Security and operational continuity	Require reasonable administrative, technical, and physical safeguards; prompt incident notification and remediation; continuity and transition provisions; and clear responsibility for service providers acting on the partner's behalf.	[Baseline: Institutionally specific.]; technical matters that need to be determined by institutional standards
Economics and reciprocal value	The partner bears its own project costs and supports the library's incremental handling, staffing, conservation, metadata, storage, and legal costs. Additional value should be proportionate where the partner receives commercial, AI-development, or strategic benefit.	[Baseline: The partner bears its own project costs and supports the library's incremental handling, staffing, conservation, metadata, storage, and legal costs.]; form and amount variable.
Non-exclusivity and competition	Avoid exclusivity except for a narrow, justified, time-limited need.	[Baseline: The ability to digitize or reuse digital content should be non-exclusive.]; narrow exceptions
Legal protections and transparency	Use knowledge-qualified library representations; avoid warranties of copyright ownership or comprehensive clearance. Allocate indemnity and liability to reflect partner technology and downstream uses. Preserve disclosure of contract terms to boards, members, counsel, auditors, and stakeholders; do not make principal terms confidential.	[Baseline: Avoid warranties of copyright ownership or comprehensive clearance; ensure indemnity covers direct and indirect costs arising from legal action.]
Term, termination, and institutional control	Include defined terms, cure periods, termination rights, transition assistance, delivery of outstanding files, return of physical materials, survival limits, name-use controls, assignment restrictions, and acceptable governing law or a neutral formulation.	[Baseline: Governing law of the library's home jurisdiction or neutral formulation]; specific terms variable.

About the Center for Research Libraries

The Center for Research Libraries (CRL) is a network of libraries that has supported original research and multidisciplinary teaching for more than 75 years. United by a shared mission, these institutions collaborate to build and steward collections that fill critical gaps, preserve long-term access, and ensure that researchers have access to the resources necessary to advance scholarship across disciplines. Researchers worldwide benefit from CRL's shared collections, which extend and enhance local library resources, and from services designed to maximize their impact and accessibility.

With a membership of approximately 200 libraries and the colleges, universities, and research communities they serve, CRL brings together ideas, experience, and emerging priorities from across the profession. As an independent, member-governed nonprofit organization, CRL provides a structure for turning that collective insight into practical action: coordinating shared initiatives, developing common approaches, and addressing issues that individual institutions cannot solve as effectively on their own.

Learn more about membership and apply to join at:

<https://www.crl.edu/crl-membership>

