MSR145

Measuring the Elephant in the Room: Quantifying Legal Barriers to AI Processing of Full-Text in Systematic Reviews

Authors: Artur Nowak[1], Marie Lane[2], Seye Abogunrin[2]

1. Evidence Prime, Krakow, Poland
2. F. Hoffmann-La Roche Ltd, Basel, Switzerland

Why This Matters

Generative AI can speed up full-text screening and data extraction—two of the most labour-intensive steps in HTA/HEOR evidence synthesis. The challenge is not only in technology but also in licensing: specifically, the copyright limitations on AI inference processing of full-text PDFs. We measured real-world coverage and outline practical, compliant operating paths.

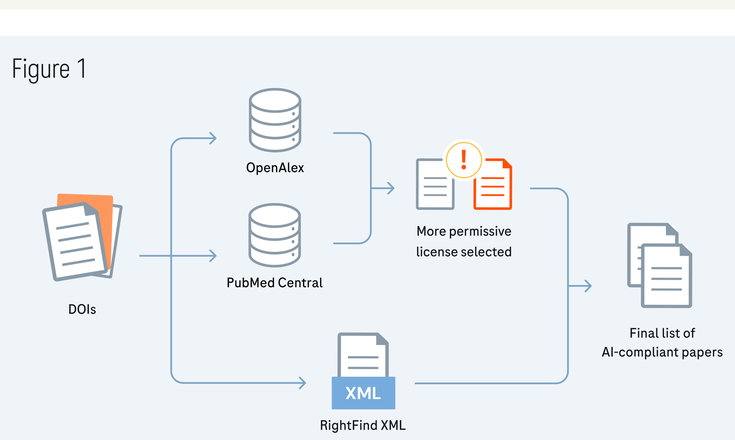
Objectives

1. Quantify how many recent review articles permit commercial text-and-data mining (TDM)/AI inference.
2. Assess agreement between open-access (OA) metadata sources.
3. Test whether a rights-vetted XML service improves coverage.
4. Provide clear operating paths and guardrails for compliant RAG-style workflows.

Methods

- Corpus: 6,336 PDFs across 49 reviews deduplicated to 3,712 unique DOIs.
- Matched to OpenAlex and PubMed Central (PMC); licence strings normalised as commercial-friendly (e.g., CC-BY/CC0/public-domain/MIT) vs restricted.
- Conflict resolution: where OpenAlex and PMC disagreed, the more permissive term was selected.
- Non-commercial (NC) clauses treated as prohibitive for AI inference.
- Checked the same DOIs against a RightFind-family XML service for AI-inference-friendly flags.

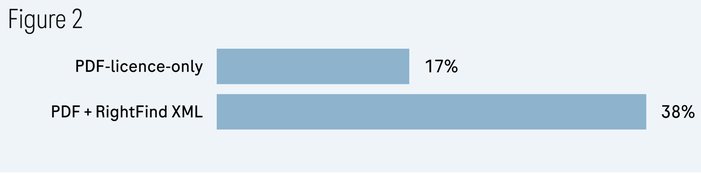
Figure 1



Results

- Licence visibility: OpenAlex had a record for every DOI, but only 1,584/3,712 (43%) exposed a licence.
- Inconsistent metadata: 26 items labelled as closed access in OpenAlex paradoxically carried OA licenses.
- Disagreements between OpenAlex and PubMed Central: Among 440 papers with dual metadata, 39 (9%) showed true licence conflicts; PMC was more permissive in 15 cases.
- Commercially usable via open access licence: 618/3,712 (17%); 3,094 remained restricted or unknown.
- Filling the gap with RightFind XML files: AI-inference-friendly files for 1,332 papers, unlocking access to 794 otherwise-restricted studies and expanding the usable corpus to ≈38%.

Figure 2



Compliant AI use

1 No separate license (OA-only path)

Operate solely on open-access content with commercial-friendly licences (e.g., CC-BY/CC0/public-domain/MIT). In our corpus, that yields ~17% coverage. The ongoing initiatives such as TDM Reservation Protocol (TDMRep) will hopefully make it easier to determine the rights in an automatic (machine-readable) fashion [1]

Operating notes
- Normalise and reconcile OA licences across sources (OpenAlex + PMC). Treat NC as prohibitive for AI inference.
- Expect gaps/inconsistencies; document your decision rules.
- Prefer structured facts over verbatim passages; keep any quotes short and within permitted uses.

2 Relying on legal exceptions without a specific licence (check opt-outs)

For pure-RAG applications (see: RAG ≠ Training), you may rely on TDM/temporary-copy exceptions. However, you must first check and honour publisher/title opt-outs. Many large publishers maintain explicit restrictions. Without a negotiated licence, coverage typically remains close to the OA-only path.

Operating notes
- Maintain an opt-out registry; enforce it at retrieval time.
- Keep processing ephemeral (no persistent embeddings/caches).
- Log provenance and who had lawful access for each processed item.

3 Using a TDM-cleared service (e.g., RightFind XML) — still not 100%

A rights-vetted XML feed can materially expand compliant coverage (here: ~38%), but it does not reach 100%. Reasons include publisher exclusions, incomplete back-files, and title-level licence carve-outs.

Operating notes
- Treat XML flags as necessary but not always sufficient; continue to enforce title-level restrictions and internal access controls.
- Apply the same ephemerality and minimal reproduction principles as above.
- Audit which items enter your AI workflow via the XML route.

4 Using a specific RAG-use–cleared licence — best coverage, with caveats

Some licences explicitly allow AI inference/RAG on lawfully accessed content. This offers the best practical coverage, but comes with non-negotiable operational requirements:

Deployment & access
- Run entirely on-premises or within your controlled infrastructure; avoid sharing PDF contents with third parties (e.g. LLM model providers).
- Enforce per-user access control inside the tool: only named users (no batch or generic accounts) can access full texts

Use scope
- Limit AI to inference (see box). No model training on licensed content.

Reproduction & outputs
- Minimise reproduction of the work: favour structured fields over verbatim text. If quotes are required, keep them short and within permitted use (e.g., internal dossiers, regulatory submissions), and visible only to authorised users.

Ephemerality
- To stay compliant, the protected content can be only processed as a “temporary copy” [2]. Use dynamic/ephemeral RAG only: short lived caches; avoid persistent embeddings of full-text and reusing the content outside of a well-defined, single task

Governance
- Maintain per-record audit trails (source, licence/flags, user identity, processing actions).
- Keep an up-to-date policy map linking licence clauses to technical controls.

RAG ≠ Training

- RAG/inference retrieves licensed content at query time with minimal persistence, aligned with licence terms and opt-outs.
- Examples of RAG use in evidence synthesis include screening, classification, tagging, summarisation, and structured data extraction.
- Training reproduces works into model parameters.

Take-home messages

- OA-only: expect ~17% usable for commercial AI use.
- Exceptions + opt-outs: coverage typically stays near OA-only levels; opt-outs must be honoured.
- TDM-cleared XML: improves to ~38% but is not universal; still requires guardrails.
- RAG-cleared licence: best coverage, contingent on on-premise deployment, per-user access controls, no training, minimal reproduction, and ephemeral RAG.
- Limitations: The above recommendations apply to systems with a considerable “human-in-the-loop” component. Fully automated literature processing pipelines, where PDF retrieval is done outside of the context of a specific task performed by an individual, need to be considered separately. Likewise, these rules may look different depending on the company's internal regulations or the laws in specific countries.

References

[1] TDM Reservation Protocol (TDMRep), Final Community Group Report. Accessed at: https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240510/

[2] Directive 2001/29/EC, Article 5(1): “Temporary copy” (temporary act of reproduction) – An act of reproduction that is (i) temporary, (ii) transient or incidental, (iii) an integral and essential part of a technological process, whose sole purpose is to enable (a) network transmission by an intermediary or (b) a lawful use of the work, and (iv) has no independent economic significance. These conditions are cumulative and interpreted strictly.