RAG Data Sourcing Strategy for Clinical Information Systems

This document outlines the data sourcing options for the PRISM clinical information system's RAG (Retrieval Augmented Generation) pipeline, including legal considerations and recommended approaches.

Table of Contents

Legal Framework: Facts vs. Expression

Copyright Fundamentals

Copyright law protects **expression**, not **facts**. This distinction is critical for understanding what can and cannot be used in a RAG system.

Protected (Expression)	Not Protected (Facts)
The specific wording of a journal article	The medical fact that "Metformin reduces A1c by ~1.5%"
The structure and arrangement of a textbook	Clinical dosing information
Original explanatory diagrams	Drug-drug interaction data
Creative selection/arrangement of facts	Disease prevalence statistics

Why Journal Subscriptions Don't Grant Extraction Rights

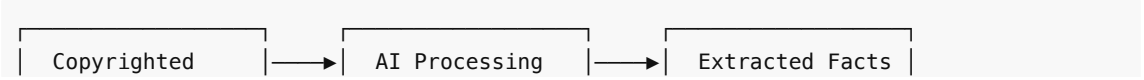
Having a personal or institutional subscription to medical journals grants **reading access only**, not:

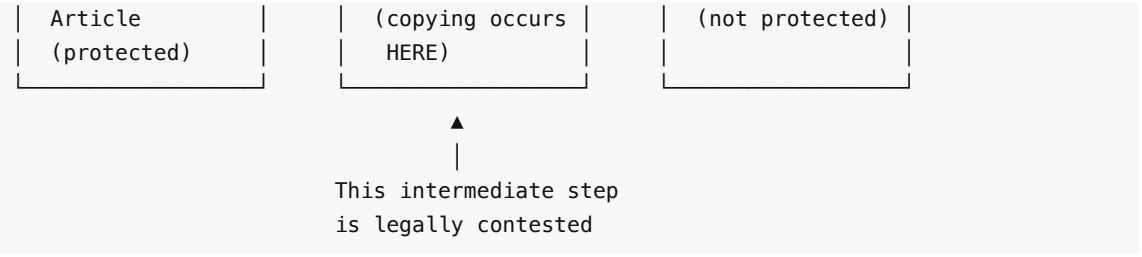
- Rights to systematically extract or scrape content
- Rights to create derivative databases
- Rights to use content in commercial products
- Rights to redistribute in any form (including as vector embeddings)

Most journal Terms of Service explicitly prohibit automated processing, creating searchable databases, and any use beyond personal research/reading.

The Intermediate Copying Problem

Even if your goal is to extract only unprotectable facts, the **process** of extraction involves copying protected expression:





Key precedent: In *Google Books v. Authors Guild*, courts allowed intermediate copying for a transformative purpose (creating a search index). However, this hasn't been tested for AI-based fact extraction specifically.

Current Legal Landscape

Legal Theory	Status	Relevance
Fair use for transformative purposes	Established, but fact-specific	May protect intermediate copying if output is truly transformative
Facts are not copyrightable	Established	Protects your use of extracted facts, not the extraction process
Database rights (EU only)	Established in EU	Protects compilations of facts in European jurisdictions
ToS/Contract violations	Separate from copyright	Even if copyright allows it, ToS may prohibit automated access
AI training on copyrighted works	Actively litigated (NYT v. OpenAI, etc.)	Outcomes will shape this area significantly

Risk Assessment for Healthcare Applications

For clinical decision support systems, unlicensed data sources create additional concerns:

1. **Regulatory risk**: FDA/medical device regulations require documented data provenance
2. **Malpractice exposure**: Unlicensed sources complicate liability questions
3. **Audit requirements**: Healthcare compliance often requires clear chain of custody for clinical content
4. **Reputational risk**: Publishers actively pursue infringement in healthcare/pharma

Recommendation: For production clinical systems, use only clearly licensed or public domain sources.

Open Data Sources

The following sources are available for use in RAG systems without licensing fees or legal ambiguity.

Tier 1: Public Domain (U.S. Government Works)

U.S. government publications are in the public domain and can be freely used.

CDC Guidelines and MMWR

- **URL**: https://www.cdc.gov/mmwr/
- **License**: Public domain ("All material in the MMWR Series is in the public domain")
- **Content**: Clinical guidelines, disease prevention recommendations, outbreak reports

- **Format**: HTML, PDF
- **API**: No official API; web scraping permitted for public domain content
- **Use case**: Infectious disease guidelines, vaccination schedules, prevention protocols

NIH Clinical Guidelines

- **URL**: https://www.nih.gov/ (various institutes)
- **License**: Generally public domain as government works
- **Content**: Treatment guidelines, clinical trial results, disease information
- **Note**: Check individual publications; some may incorporate third-party copyrighted material

AHRQ Evidence Reports

- **URL**: https://www.ahrq.gov/
- **License**: Public domain
- **Content**: Comparative effectiveness reviews, evidence syntheses, clinical decision support resources
- **Use case**: Evidence-based treatment comparisons

USPSTF Recommendations

- **URL**: https://www.uspreventiveservicestaskforce.org/
- **License**: Public domain
- **Content**: Preventive care recommendations with grade levels (A, B, C, D, I)
- **Use case**: Screening and prevention protocols

Tier 2: Open Access with Permissive Licenses

PubMed Central Open Access Subset

- **URL**: https://pmc.ncbi.nlm.nih.gov/tools/openftlist/
- **License**: Various Creative Commons licenses (check each article)
- **Content**: ~4 million full-text biomedical articles
- **APIs Available**:
  - OAI-PMH API: https://pmc.ncbi.nlm.nih.gov/tools/oai/
  - BioC API: Full text in structured format
  - E-utilities: Programmatic access
- **Important**: License varies by article. Filter for CC-BY or CC-BY-SA for maximum flexibility.
- **Cloud Access**: Available on AWS S3 without authentication

```
Recommended API workflow:
1. Use OAI-PMH to get article metadata including license
2. Filter for articles with CC-BY, CC-BY-SA, or CC0 licenses
3. Use BioC API to retrieve full text in structured format
4. Store license information with each article in your system
```

ClinicalTrials.gov

- **URL**: https://clinicaltrials.gov/
- **API**: https://clinicaltrials.gov/data-api/api (REST API v2.0)
- **License**: Public domain
- **Content**: Registry of 400,000+ clinical studies worldwide
- **Use case**: Treatment protocols, eligibility criteria, outcome measures

Tier 3: Open Medical Terminologies and Ontologies

RxNorm (Drug Terminology)

- **URL**: https://www.nlm.nih.gov/research/umls/rxnorm/
- **License**: Free, requires UMLS license (no fee, registration required)
- **Content**: Normalized drug names, ingredients, strengths, dose forms
- **API**: RxNorm API available
- **Use case**: Drug identification, ingredient lookup, dosing information

LOINC (Laboratory/Clinical Observations)

- **URL**: https://loinc.org/
- **License**: Free for use (requires accepting license terms)
- **Content**: Standard codes for lab tests, clinical observations, surveys
- **Use case**: Lab result interpretation, clinical documentation

SNOMED CT

- **URL**: https://www.snomed.org/
- **License**: Free in IHTSDO member countries (including U.S.) via UMLS
- **Content**: Comprehensive clinical terminology (350,000+ concepts)
- **Use case**: Clinical concept standardization, diagnosis coding

ICD-10 / ICD-11

- **URL**: https://www.who.int/standards/classifications/classification-of-diseases
- **License**: Free for use
- **Content**: Disease classification codes
- **Use case**: Diagnosis classification, billing codes

Tier 4: Open Access Drug and Safety Data

OpenFDA

- **URL**: https://open.fda.gov/apis/
- **License**: Public domain
- **Endpoints**:
  - `/drug/label` - Structured product labeling (67,000+ drugs)
  - `/drug/event` - Adverse event reports (4.9M+ reports)
  - `/drug/ndc` - National Drug Code directory
- **Format**: JSON REST API
- **Use case**: Drug information, safety alerts, labeling data

```
Example API call:
GET https://api.fda.gov/drug/label.json?search=brand_name:"metformin"
```

DailyMed

- **URL**: https://dailymed.nlm.nih.gov/
- **License**: Public domain
- **Content**: FDA-approved drug labeling
- **Use case**: Prescribing information, package inserts

Licensed Data Sources

For more comprehensive clinical decision support, the following require commercial licenses but provide high-quality, curated content.

Clinical Decision Support Databases

Source	Content	Licensing Model
UpToDate	Physician-authored clinical recommendations	Institutional subscription
DynaMed	Evidence-based clinical reference	Institutional subscription
Clinical Key	Elsevier clinical content	Institutional subscription
Lexicomp	Drug information database	Per-seat or enterprise

Text Mining Licenses

Major publishers offer text-mining APIs for commercial use:

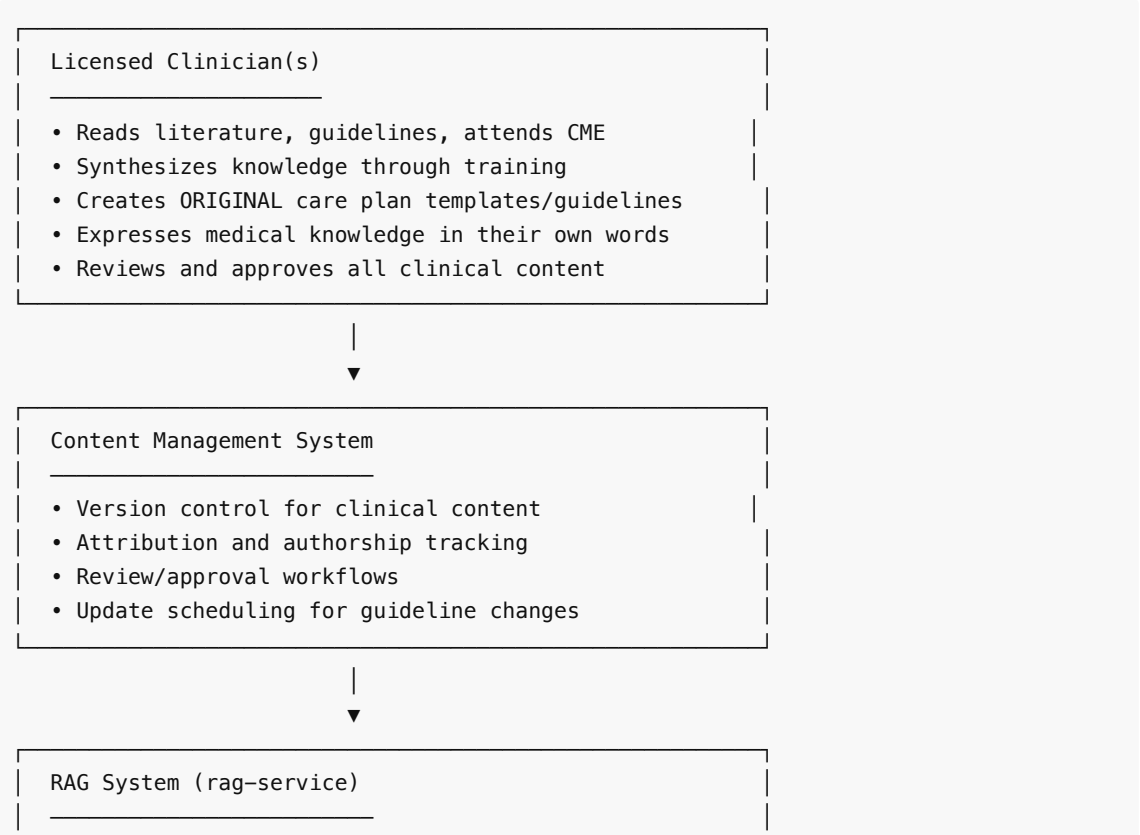
- **Elsevier**: Text Mining API (requires negotiated license)
- **Springer Nature**: TDM (Text and Data Mining) program
- **Wiley**: TDM license available

These are expensive but provide legal clarity for systematic extraction.

Clinician-Authored Content Strategy

The most legally unambiguous approach is having licensed clinicians create original content based on their training and expertise.

How It Works



```
│  • Indexes clinician-authored content                    │
│  • Stores vector embeddings                               │
│  • Retrieves relevant guidance at query time              │
│  • No copyrighted third-party text in system              │
└──────────────────────────────────────────────────────────┘
```

Legal Basis

This approach is legally clear because:

1. **Original expression**: The clinician creates new expression, not copying existing text
2. **Facts are not copyrightable**: The underlying medical knowledge is not owned by anyone
3. **Professional expertise**: Clinicians routinely synthesize knowledge as part of their profession
4. **Clear ownership**: Your organization owns the work product

Implementation Model

Option A: In-House Clinical Team

- Hire physician editors/medical directors
- Create content authoring guidelines
- Establish review and approval processes
- Budget for ongoing updates as guidelines change

Option B: Contract with Medical Writers

- Engage board-certified physicians as contractors
- Define scope and content needs
- Ensure work-for-hire agreements transfer IP rights
- Maintain editorial oversight

Option C: Hybrid Approach

- Use open sources (Tier 1-4 above) as foundation
- Have clinicians review, augment, and customize
- Clinicians add institution-specific protocols
- Original clinician content fills gaps

Content Types for Clinician Authoring

Content Type	Example	Priority
Care plan templates	"Type 2 DM Initial Management"	High
Clinical decision trees	"Chest Pain Triage Algorithm"	High
Drug selection guidance	"First-line Antihypertensives by Comorbidity"	High
Patient education	"Understanding Your A1c Results"	Medium
Protocol summaries	"Sepsis Bundle Checklist"	High
Specialty consult criteria	"When to Refer to Cardiology"	Medium

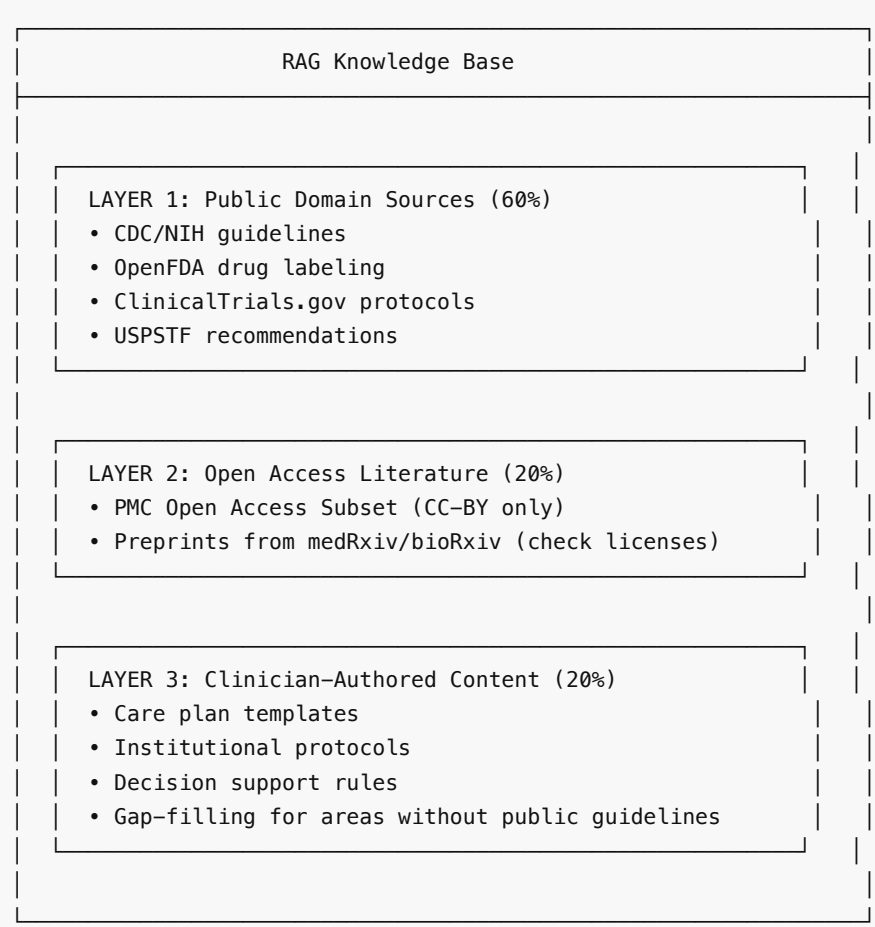
Quality Assurance

1. **Peer review**: All content reviewed by second clinician
2. **Evidence linking**: Reference source guidelines (CDC, USPSTF, etc.) without copying text

3. **Currency tracking**: Flag content for review when source guidelines update
4. **Conflict of interest**: Document author COI disclosures

Implementation Recommendations

Recommended Data Source Mix for PRISM



Integration with PRISM Services

rag-service Integration Points

```
// Suggested data source configuration
interface DataSourceConfig {
  // Public domain sources - unrestricted use
  publicDomain: {
    cdcGuidelines: boolean;
    openFDA: boolean;
    clinicalTrials: boolean;
    uspstf: boolean;
  };

  // Open access with license tracking
```

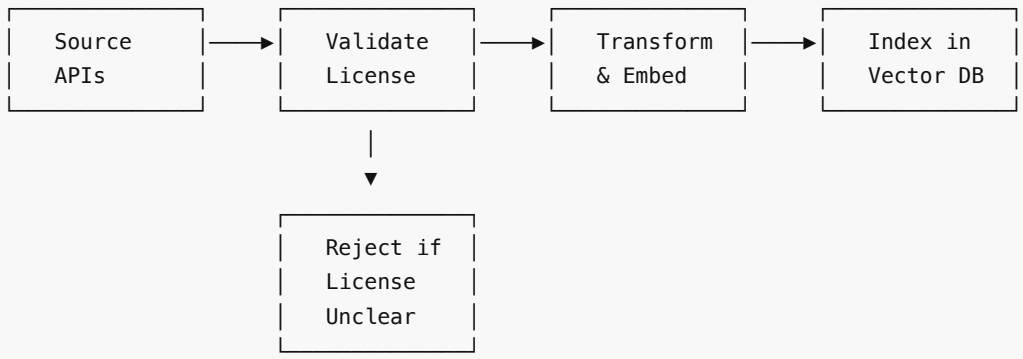
```
  openAccess: {
    pmcOpenAccess: boolean;
    allowedLicenses: string[]; // ['CC-BY', 'CC-BY-SA', 'CC0']
  };

  // Clinician-authored content
  clinicianAuthored: {
    carePlanTemplates: boolean;
    institutionalProtocols: boolean;
  };
}
```

careplan-service Integration

The `careplan-service` should pull from:

1. Clinician-authored care plan templates (primary source)
2. Relevant CDC/NIH guidelines for evidence basis
3. OpenFDA for drug-specific information

Data Ingestion Pipeline



Metadata Requirements

For each document in the RAG system, store:

```
interface DocumentMetadata {
  // Source tracking
  sourceType: 'public_domain' | 'open_access' | 'clinician_authored' | 'licensed';
  sourceUrl: string;
  retrievalDate: Date;

  // License information
  license: string;           // e.g., 'public_domain', 'CC-BY-4.0'
  licenseUrl?: string;

  // For clinician-authored content
  author?: {
    name: string;
    credentials: string;     // e.g., 'MD, FACP'
```

```
    institution: string;
  };
  reviewedBy?: string;
  approvalDate?: Date;

  // Currency tracking
  contentDate: Date;          // When the source content was published
  nextReviewDate?: Date;      // When to check for updates

  // Provenance for auditing
  ingestionPipeline: string;
  version: string;
}
```

Summary

Approach	Legal Clarity	Cost	Content Quality	Recommendation
Public domain (CDC, NIH, FDA)	Unambiguous	Free	High (authoritative)	Primary source
PMC Open Access (CC-BY)	Clear with license tracking	Free	Variable	Secondary source
Clinician-authored	Unambiguous (you own it)	Moderate-High	Customizable	Fill gaps, customize
Licensed databases	Clear (contractual)	High	Very high	Consider for scale
Journal scraping/extraction	Legally contested	Free	High	Avoid for production

Next Steps

1. **Immediate**: Set up ingestion pipelines for public domain sources (CDC, OpenFDA, ClinicalTrials.gov)
2. **Short-term**: Implement PMC Open Access ingestion with license filtering
3. **Medium-term**: Develop clinician content authoring workflow
4. **Ongoing**: Monitor legal developments in AI and copyright

---

References and Resources

APIs and Data Sources

- PubMed Central Open Access Subset
- PMC Text Mining Resources
- OpenFDA API Documentation
- ClinicalTrials.gov API
- CDC MMWR
- USPSTF Recommendations

- NLM SNOMED CT Licensing

Legal Background

- NCBI Copyright Information
- Google Books v. Authors Guild (fair use precedent)
- NYT v. OpenAI (ongoing litigation on AI training)

Document created: December 2024 Review scheduled: Quarterly or upon significant legal developments