



**THE BROWN INSTITUTE
FOR MEDIA INNOVATION**

History Lab AI: Making Historical Archives Conversational

An Experiment in AI-Powered Historical Research

Columbia University History Lab

Meet the Team



Matthew Connelly

Professor of History, Columbia University



Raymond Hicks

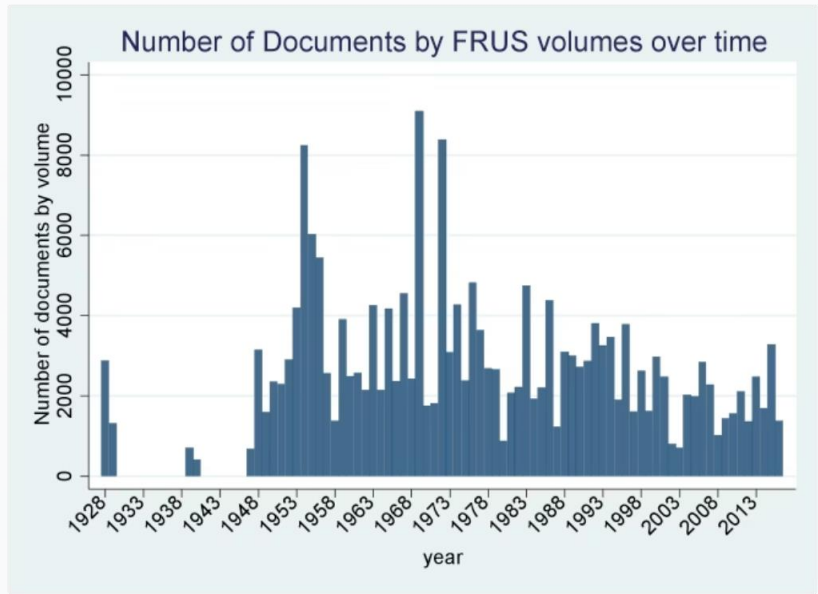
Project Manager, History Lab



Nick Chimicles

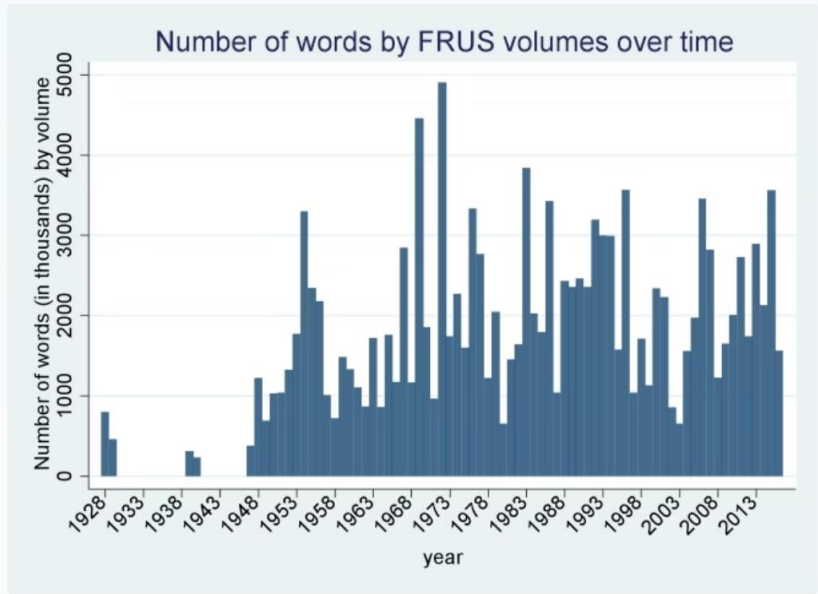
Project Lead, History Lab AI

Declining Releases



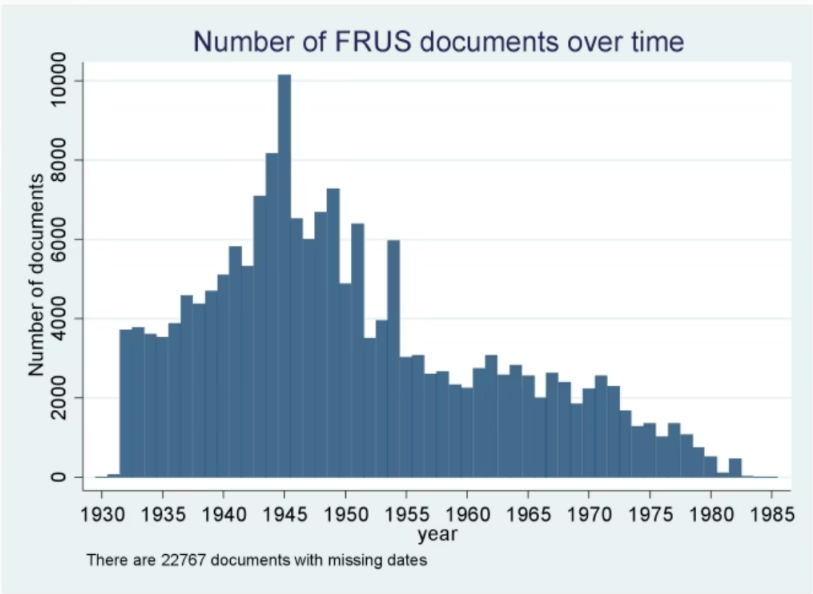
Panel A: Fewer Documents in FRUS

Foreign Relations of the United States volumes contain fewer individual documents over time



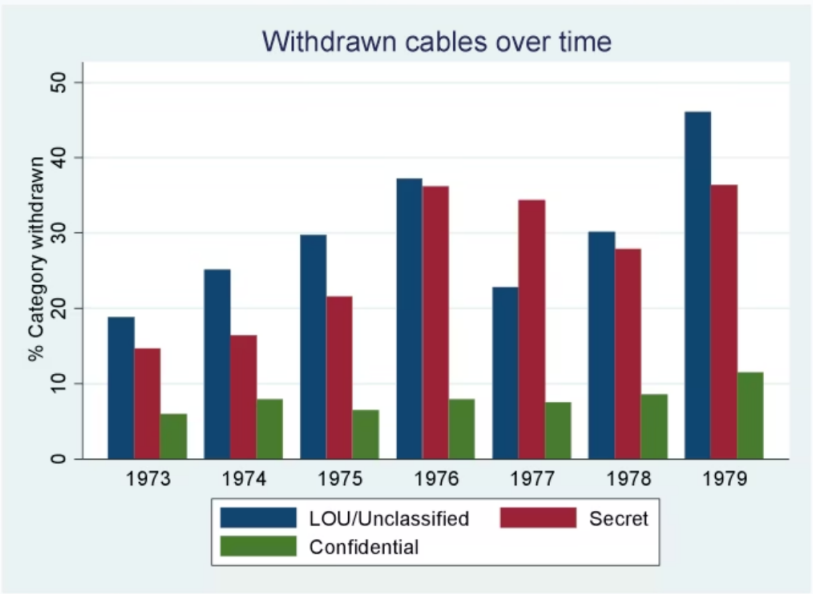
Panel B: Longer Documents Compensate

Average document length increases to maintain volume size



Panel C: Sharp Drop After 1960s

Dramatic reduction in document releases following the Cold War era

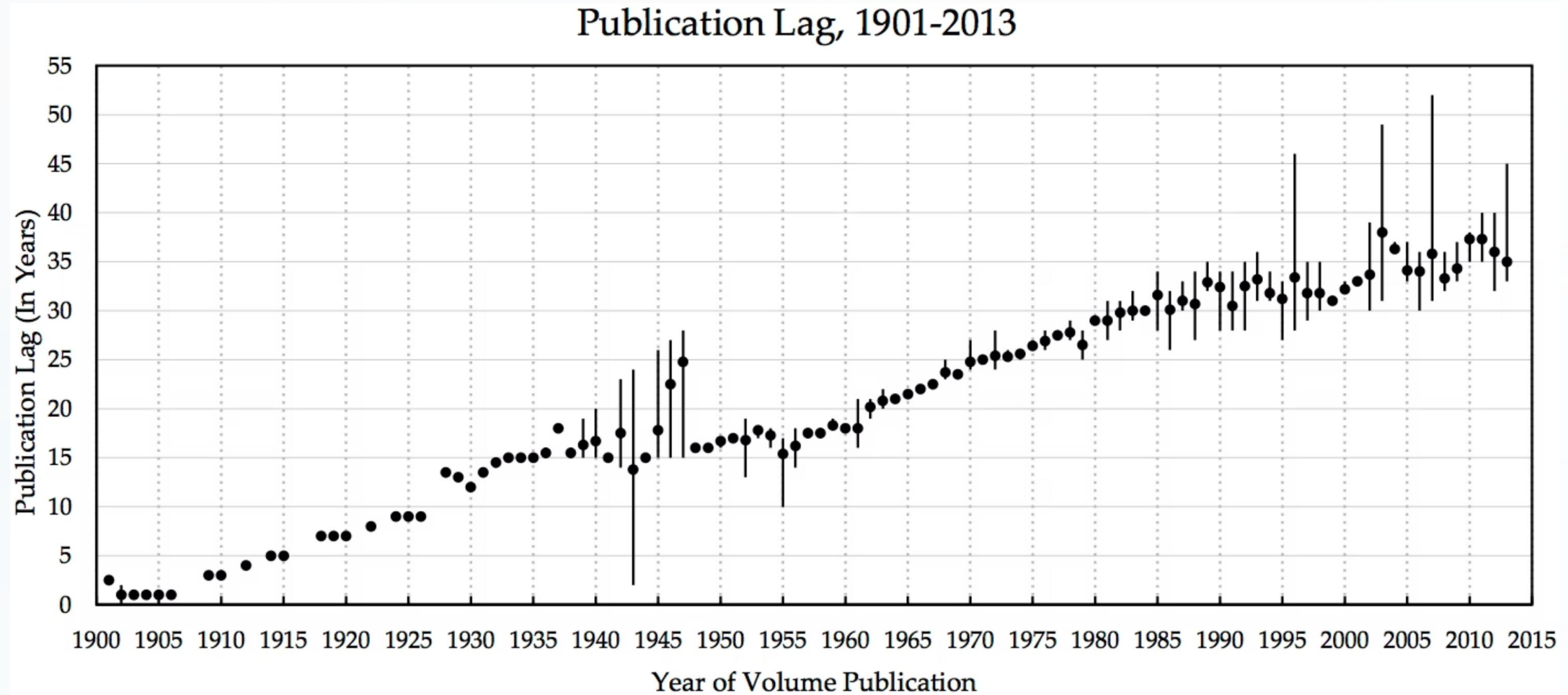


Panel D: No Top Secret Cables

Highest classification levels remain completely withheld

The Declassification Crisis

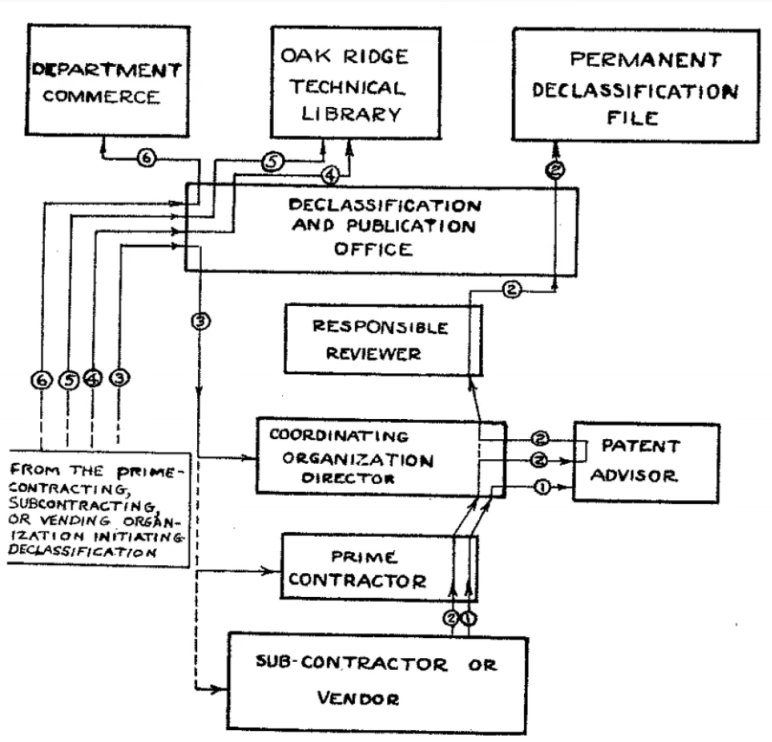
Publication Lag Increasing Dramatically



The marker displayed for each year corresponds to the *average* publication lag of all the volumes released during that year. The bars above and below each marker show the range of the *lowest* and *highest* publication lag of the volumes released during that year.

Process Comparison

How Declassification Has Changed



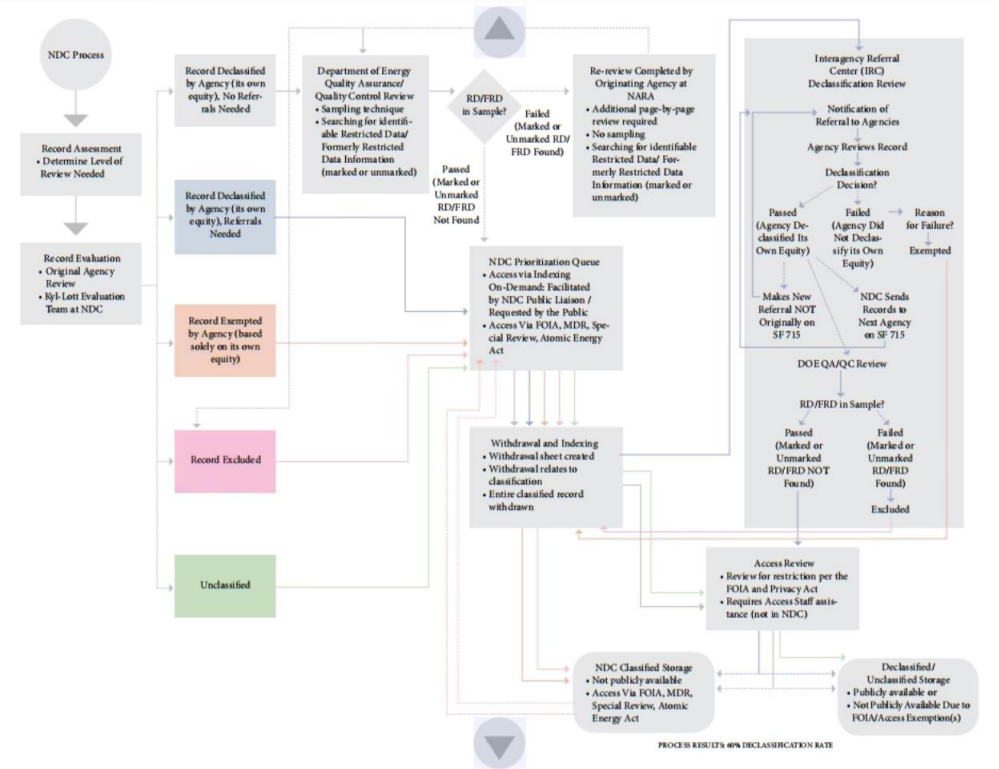
1946: Simple Flow Chart

Straightforward review process with clear decision points and minimal bureaucratic layers

01

Multiple Agency Reviews

Documents must pass through numerous government departments



2015: Complex Maze

Intricate web of reviews, committees, and approval processes

02

Interagency Referral Center

Central coordination point that often creates bottlenecks

03

Years of Bureaucratic Delays

Process can take decades from initial request to final release

What is History Lab and what does it do?



World's largest database of declassified records

Spanning decades of U.S. foreign, military, and intelligence history



Applied research for government legal mandates

Support transparency and accountability through data-driven analysis



Turns documents into data

Transform historical archives into searchable, analyzable datasets



Develops tools to explore history

Create innovative digital tools for historical research and discovery

The Declassification Engine



Using AI to Understand Classification

Machine learning models trained on patterns in classification decisions



Train on Known Classified/Unclassified

Learning from historical examples of what gets released vs. withheld



93% Accuracy Predicting Classification

High precision in identifying what should remain classified



Identify Over-Classification Patterns

Spotting documents that are unnecessarily kept secret



Make Process Transparent

Providing clear rationale for classification decisions



The FOIArchive

5M+

Declassified Documents

Comprehensive collection spanning
decades

18M+

Pages

Millions of pages of historical content

Collections Include:

- CIA CREST (1941-2005)
- State Department cables (1973-1979)
- FRUS volumes
- Presidential Daily Briefings (1946-1977)

The Original FOIA Document Challenge

No OCR

Millions of scanned images, not searchable text

Scattered Silos

CIA CREST here, State cables there, FRUS elsewhere

No Standards

Different formats, metadata, download methods per agency

Zero Context

No entity linking, topic grouping, or relationship mapping

Minimal Metadata

Basic dates, titles, and descriptions at best

The Problem: 5 million documents locked away in unusable formats across disconnected government sites



According to the finding aid,
the needle is filed under
'Misc.'

History Lab's Foundation - The FOIArchive



Unified 5M+ Documents

All collections in one searchable database



High-Quality OCR Processing

Every document fully text-searchable



Standardized Metadata

Consistent structure across all sources



Natural Language Processing

Topic modeling, Named Entity Recognition

The Evolution of History Lab Search:

Advanced Filtered Search

History Lab AI elevates archival research with intelligent filtering and contextual understanding.



Smart Entity Recognition

Automatically links diverse mentions (e.g., "Castro" = "Fidel Castro" = "Cuban Leader") to the same historical figure or organization, ensuring no relevant document is missed.



Temporal Distribution Visualization

Explore document clusters through an interactive timeline view, enabling precise date filtering while preserving crucial historical context.



Rich Metadata Enrichment

Leverages detailed metadata such as classification levels, document types, and source collections for highly granular and accurate search queries.

The Evolution of History Lab Search:

Topic Modeling Search

Moving beyond simple keywords, FOIArchive leverages advanced machine learning to uncover deep thematic insights within the archives.



Machine Learning Topic Discovery

Automatically identifies prevalent themes and subjects across millions of documents, effectively cutting through bureaucratic phrasing and jargon.



Semantic Clustering

Groups related documents and ideas based on their underlying meaning, rather than exact keyword matches. For example, a cluster might reveal documents pertaining to "agenc, intellig, oper, activ, group" to broadly define intelligence operations.



Browse by Concept, Not Keywords

Enables researchers to explore documents by conceptual meaning, fostering the discovery of unexpected connections and hidden historical narratives.

The Evolution of History Lab Search: Interfaces

FOIArchive Search

First time here? Learn more about our collections [here](#) and watch this brief [screencast](#) to see how to use the search interface. More specific queries returning < 2000 documents run faster and return metadata and text.

Full-Text

Enter search terms

Corpus

Originally restrict to specific corpora

Original Classification

Originally restrict to specific classificati...

People, Places, Organizations...

Originally restrict to specific entities

Date Range

YYYY/MM/DD – YYYY/MM/DD

☒ All entities appear in document

☒ Include documents without a date

Search

Query the FOIArchive via topics derived by topic modeling. You can find more information about topic modeling [here](#).

Corpus		Topic			
frus		agenc, intellig, oper, activ, group			
Score	Document	Title	Date	Corpus	Classif
0.70	View	263. Office of Special Operations Directive No. 18/5	1948-03-29	frus	top se
0.68	View	255. Memorandum from President Kennedy to McCone, January 16	1962-01-16	frus	unknc
0.65	View	260. Office of Special Operations Directive No. 18/5 (Interim)	1948-02-24	frus	secret
0.65	View	358. Report From the Intelligence Survey Group to the National Sec	1949-01-01	frus	top se
0.63	View	427. National Security Council Intelligence Directive No. 7	1948-02-12	frus	secret
0.62	View	160. National Intelligence Authority Directive No. 5	1946-07-08	frus	top se
0.62	View	114. Memorandum by the Director of Central Intelligence's Executiv	1946-07-11	frus	secret
0.62	View	423. National Security Council Intelligence Directive No. 5	1947-12-12	frus	top se
0.62	View	103. Memorandum of Agreement on Direction and Supervision of U.	1966-08-10	frus	unknc
0.62	View	11. Memorandum From the Chairman of the U.S. Intelligence Board	1973-10-03	frus	confid

The Evolution of History Lab Search:

AI Conversational Search

Engage with history like never before, using an intuitive conversational interface powered by advanced AI.



Natural Language Queries

Ask questions in plain English and engage in a dynamic dialogue with historical archives.



Semantic Understanding

Vector search incorporates meaning across all terminology variations, transcending keyword limitations.



Contextual Synthesis

The AI reads and summarizes information across multiple documents to provide comprehensive, integrated answers.



The Evolution Continues

Building on our robust search infrastructure, History Lab AI introduces cutting-edge conversational capabilities.



ASSISTANT


06:38 PM HISTORYLAB

I will search for intelligence assessments of Soviet capabilities and intentions leading up to the December 1979 invasion of Afghanistan.

I will start with a broad search covering 1978-1979, then narrow the timeframe if needed. I will focus on documents from the

Live Demo

Live Interface Demonstration

 history-lab.ramus.network

History Lab



Usage Statistics

~300

Users

Pilot program participants

800

Conversations

Research sessions conducted

3.6K+

Messages

User interactions with AI

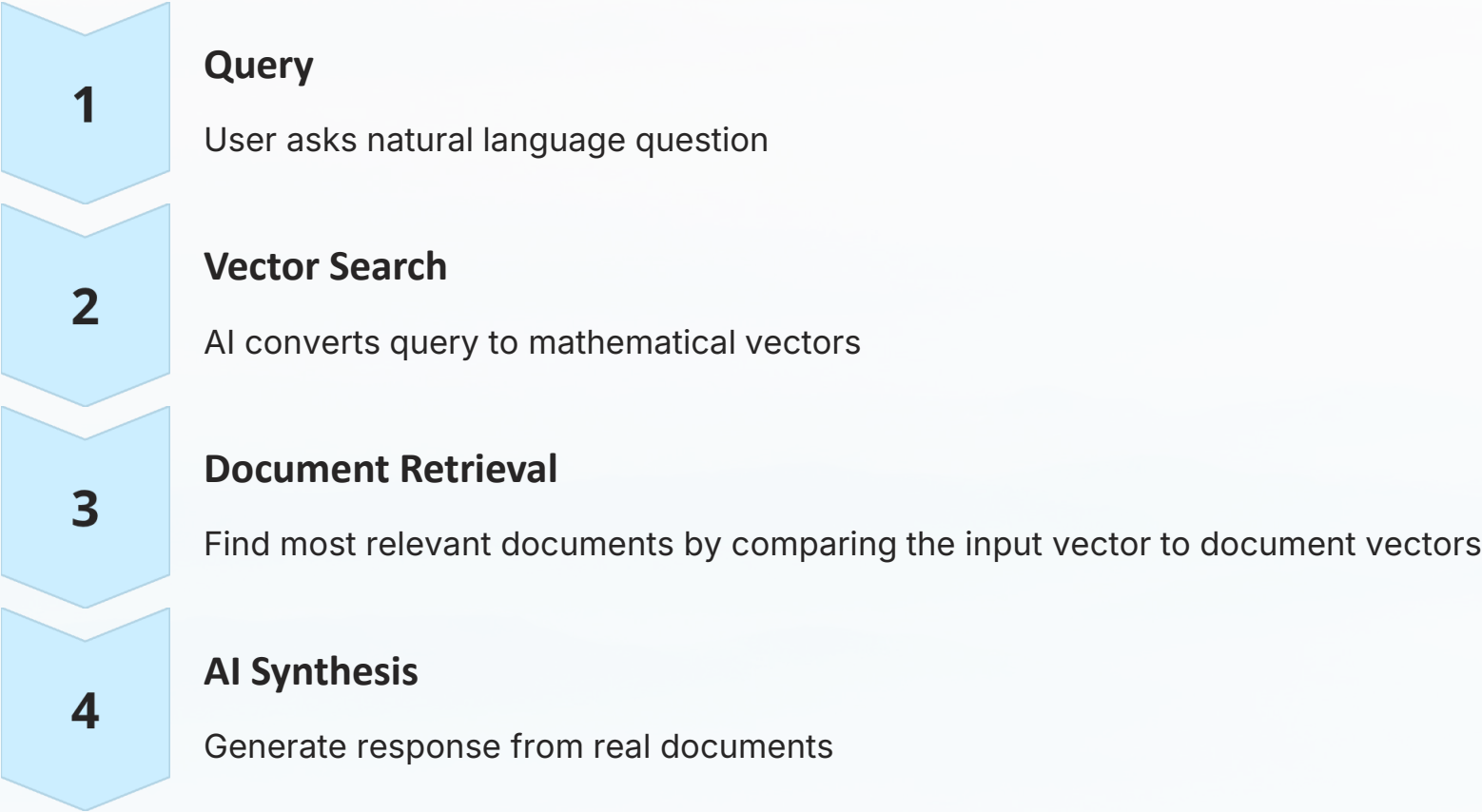
2.3K+

Tool Calls

AI document retrievals

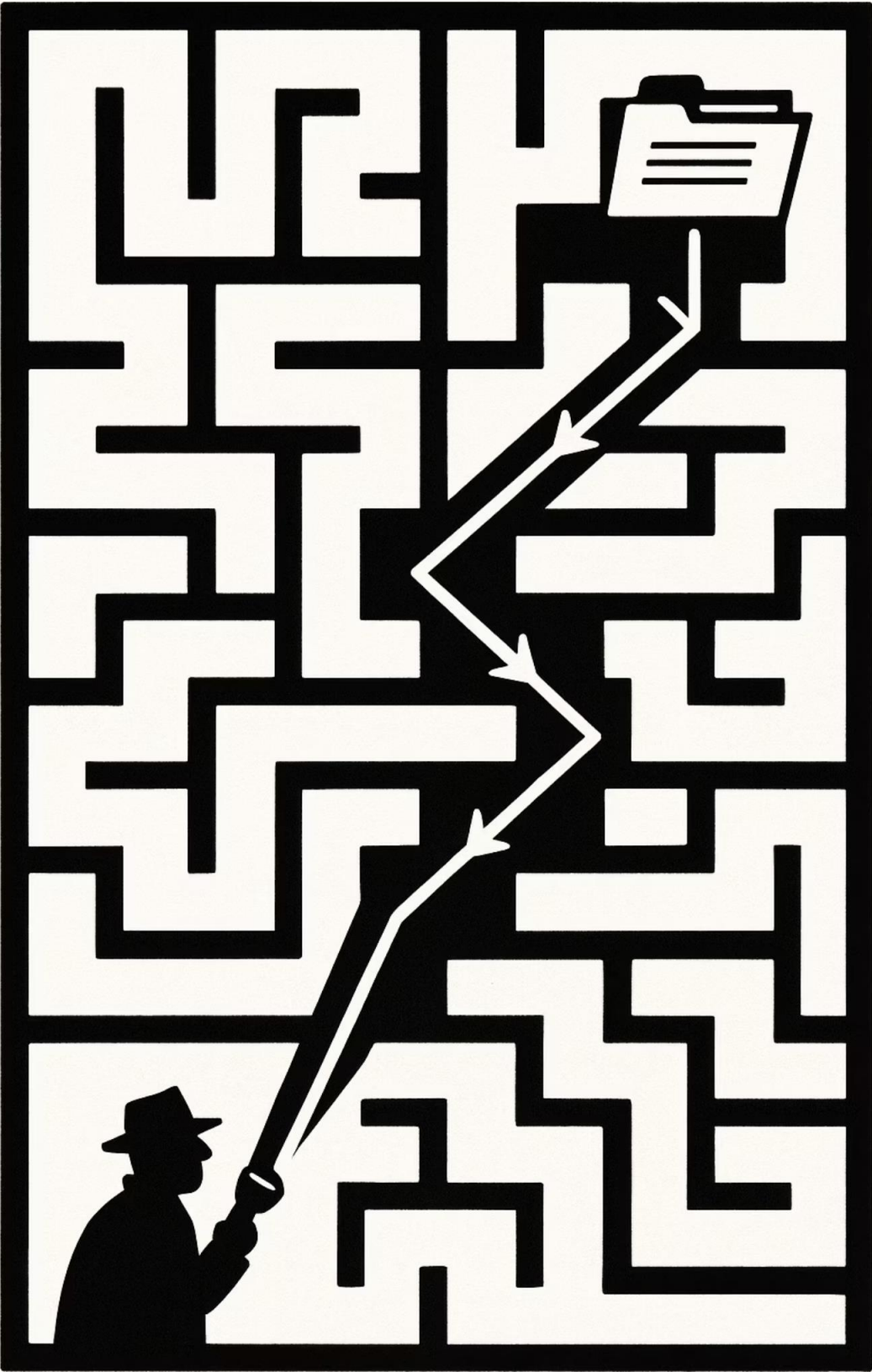
How It Works - RAG

Retrieval Augmented Generation (RAG)



✔ Reduced hallucination by only citing real documents

One AI model + vector database working together to provide accurate, source-backed answers



Effective Querying

Be Specific with Dates and Actors

Include timeframes and key people to narrow results effectively

Break Complex Queries into Parts

Divide complicated questions into manageable components

Use Document Language

Match the formal, bureaucratic tone of government documents

More Context is Better

Provide additional contextual words rather than fewer

Iterate and Refine

Build on previous queries to drill down into specific topics



Good: "CIA assessment of Soviet threat"



Bad: "CIA feelings about Soviet union"

Semantic Search

Vectors Represent Ideas Mathematically

Traditional Search

Traditional search looks for exact word matches. It processes queries literally, only returning results that contain the precise keywords used.

Example 1: "Boats"

- ❗ Traditional search for **"Boats"** might only find documents with that exact word.

Semantic search for **"Boats"** also finds results for *"ships," "vessels,"* and *"naval craft"* because it understands their related meaning.

Semantic Search

Semantic search understands meaning and context, not just keywords. It interprets the intent behind a query to provide more relevant results.

Example 2: "Glasnost"

- ✅ Traditional search for **"Glasnost"** would only return documents explicitly mentioning "glasnost."

Semantic search for **"Glasnost"** intelligently finds related concepts like *"openness," "political liberalization,"* or *"Gorbachev reforms,"* understanding the broader context.

Cosine similarity is a key technique used in semantic search to find nearest vectors in a multidimensional space, effectively connecting related concepts even when exact words don't match.

Generational Divide

Younger Users



Conversational Queries

Ask complex, multi-part questions naturally



Talk to AI Like Colleague

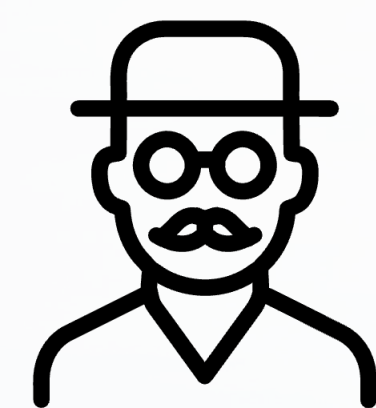
Comfortable with AI as research partner



Voice Dictation

Prefer speaking queries over typing

Older Historians



Single Keywords

Prefer simple, direct search terms



Just Names

Focus on proper nouns and specific entities



Traditional Search Mindset

Approach AI like a library catalog

It's important to note that this observed generational divide is based solely on user data collected during our pilot program and is not intended as a universal statement.

What the AI has been prompted to do

01	02	03
Input Filtering	Query Creation	Document Retrieval
Corrects spelling and clarifies user queries for optimal processing	Generates appropriate search queries tailored to the document corpus	Searches through millions of documents to find relevant matches
04	05	
Relationship Mapping	Corpus Summarization	
Identifies connections and patterns across retrieved documents	Synthesizes findings from large document collections	

Chunking Strategy

What Makes Your Documents Most Discoverable?

Embedding Decisions

- Is descriptive metadata enough? Can work if descriptions are rich enough
- Full document embedding? Needed when metadata is unreliable
- Our approach: Both metadata and document body (because descriptive metadata was unreliable)
- Do light testing to find what works best for your collection

Advanced Chunking Approaches

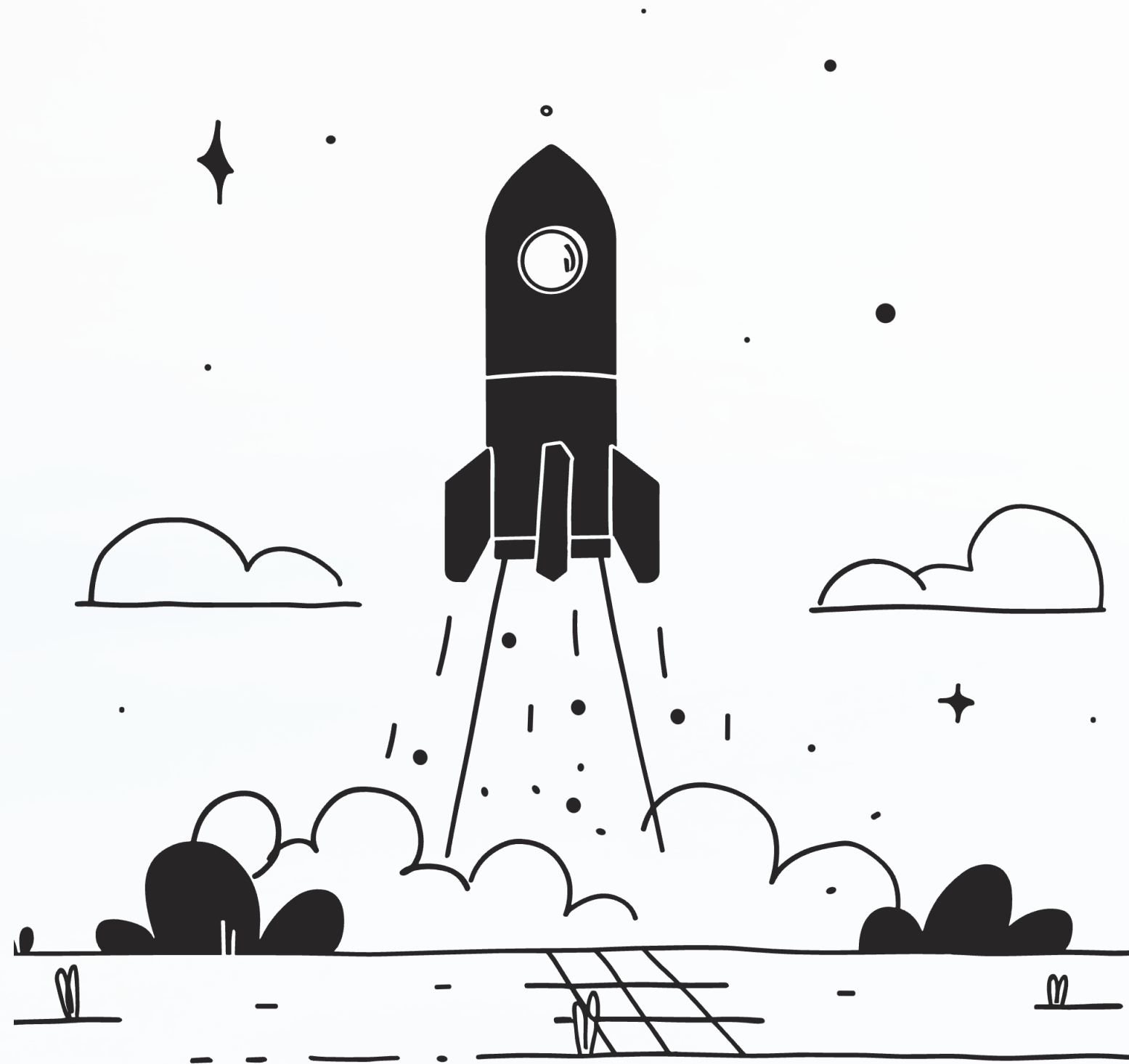
- 512 tokens per chunk (~750 words) - embedding model limitation
- Simple approach: Fixed-size chunks with overlap
- Advanced: Chunk by document structure (sections, tables) for born-digital documents
- Add overlap between chunks to preserve context at boundaries

Key Considerations

- Balance between context and precision
- OCR quality impacts embedding effectiveness
- Test different approaches - what works varies by collection
- Token limits force difficult decisions about what to include



Implications of the new new semantic search paradigm



Impact on Research



Democratizing Primary Sources

Making historical documents more accessible to researchers, no need for keyword expertise



Cross-Collection Discoveries

Finding connections between documents across different archives and time periods based on their contextual meanings



Questions That Took Days → Minutes

Dramatically reducing research time from days of manual searching to minutes of AI-assisted discovery



New Research Possibilities

Opening entirely new avenues of inquiry that were previously impractical or impossible

This significant impact would only be realized if users are properly educated in how to use these tools effectively.

Implications for Archives - Meeting Users in the AI Era



The Discovery Paradigm Shift

Rich Items at Scale

- Past wisdom: Collection-level descriptions sufficient for discovery
- New reality: AI thrives on item-level detail and full-text OCR
- Users now expect to find specific documents, not just collections



Critical Infrastructure for AI Discovery

- Quality OCR: Foundation for everything – both search and AI-generated metadata
- Rich Descriptions: Natural language descriptions matter more than controlled vocabularies
- Basic Structure: Date, creator, subject still useful, but extensive keyword tagging less critical
- Scale Priority: Better to have 10,000 documents with good OCR than 100 perfectly catalogued



The Conversational Shift

- Users ask "What did the CIA know about Cuba in 1962?" not search "Castro AND intelligence AND assessment"
- AI translates natural questions into semantic searches across your full text
- Rich prose descriptions become more valuable than fielded metadata



Opportunity: Let AI Help

- Use AI to generate initial item descriptions from good OCR
- Focus human expertise on verification and context
- Prioritize digitization and OCR quality over perfect metadata

The MCP Future - Archives in the AI Ecosystem

Model Context Protocol (MCP): Your Gateway to AI

Open standard that lets **any** AI system query your collections while **you** control access



 Humane Ingenuity

AI and Libraries, Archives, and Museums, Loosely Couple

A new framework provides a way for cultural heritage institutions to take advantage of the technology with fewer misgivings, and to serve students, scholars, and the public better



1

How It Works

- Your archive runs an MCP server (like having an API)
- AI assistants connect to request information
- You define what's accessible: just metadata, full text, or custom views
- Users get authoritative answers with proper citations back to your site

2

Already a Reality

- Integrates with Claude, ChatGPT, Gemini, and open-source models
- Each institution maintains its own MCP server
- Great potential for network effects in the cultural heritage space
- "Loosely coupled" - you keep control, AI gets better answers

Building Your Own

Open Source Code Available

Complete codebase and documentation available for researchers and institutions wanting to build their own AI-powered archives.

Key Technical Decisions:

- **Embedding Metadata and/or Text**

All about finding the right balance. Depending on your collection metadata may be sufficient

- **Embedding Model (512 tokens)**

AI model for converting text to vectors (Google offers free embedding API ... for now)

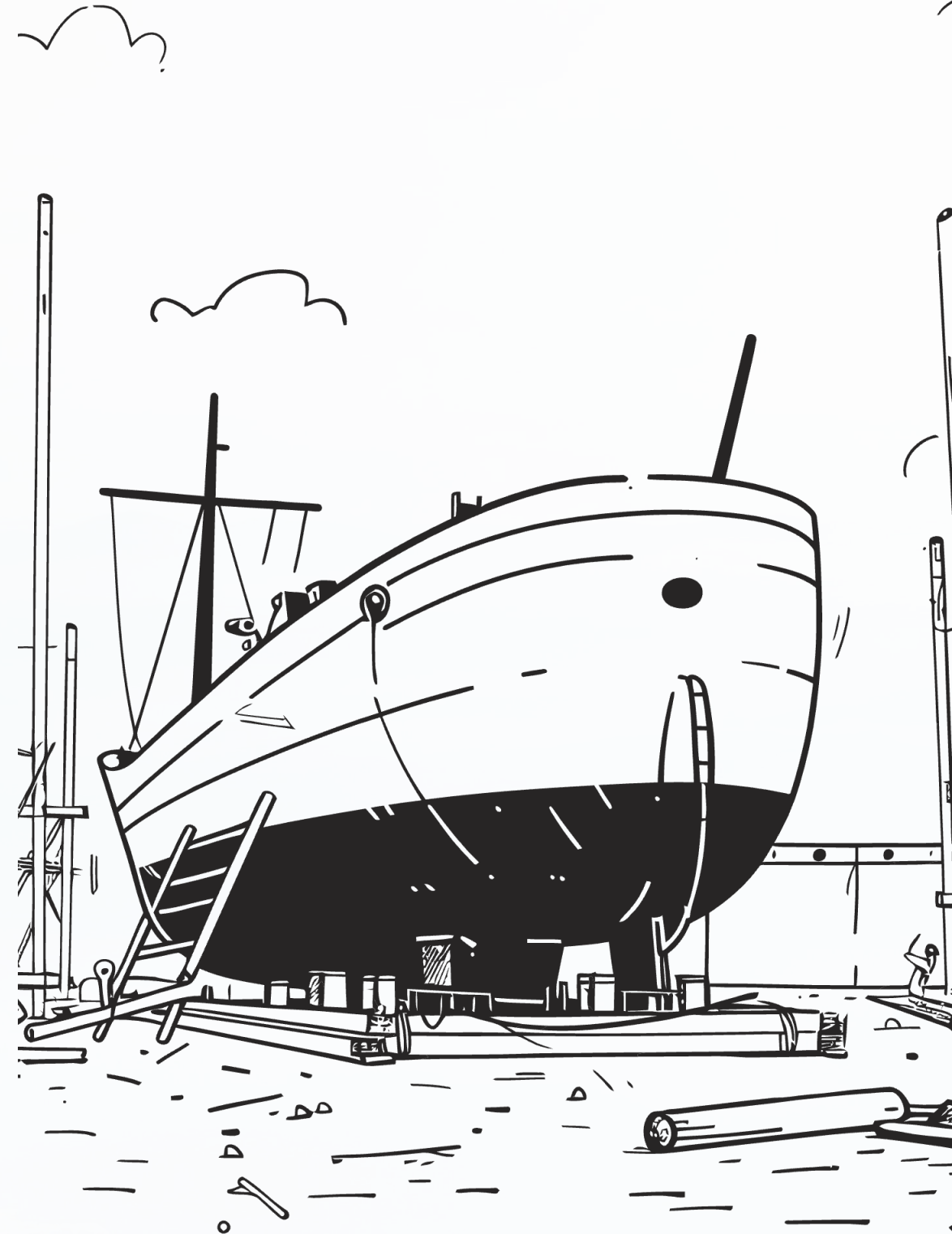
- **Chunking Strategy**

How to divide documents for optimal search (overlap, structure etc.)

- **Vector Database Choice**

Storage and retrieval system selection (Suggestion: Pinecone)

GitHub: <https://github.com/Ramus-Network/chat-agent-historylab>



Contact & Resources



Website

lab.history.columbia.edu



Research Tool

<https://lab.history.columbia.edu/history-lab-llm>



Email

info@history-lab.org



History Lab AI GitHub

<https://github.com/Ramus-Network/chat-agent-historylab>



History Lab Github

<https://github.com/history-lab>

Your Feedback Matters!

We value your input to help us improve History Lab AI and develop future features. Please take a few moments to share your thoughts by completing our survey.

History Lab Conversational Search Interface - Feedback Form

Please provide your feedback after trying the History Lab tool at: <https://lab.history.columbia.edu/history-lab-llm>

What is your role/title?

☐ Librarian


☐ Archivist

☐ Research Support Staff

☐ Faculty/Researcher


☐ Graduate Student

☐ Other

 Google Docs

History Lab Conversational Search Interface - Feedback Form

Please provide your feedback after trying the History Lab tool at: <https://history-lab.ramus.network/>



Questions & Discussion

Q&A

We welcome your questions about History Lab AI, the declassification process, and the future of AI-powered historical research.

Website

<https://lab.history.columbia.edu>

Research Tool

<https://lab.history.columbia.edu/history-lab-llm>

Contact

info@history-lab.org

Survey

[Feedback Form](#)

