ITDM Blog

Beyond OCR: How Multimodal AI Unlocks Historical Archive Value

The Strategic Gap in Newspaper Digitization

Across the publishing industry, major initiatives have focused on digitizing historical newspaper archives, yet most still fail to unlock their full strategic potential. A prominent U.S. newspaper illustrates this challenge: decades of print editions have been scanned and made available to subscribers online, but the resulting archives remain difficult to search with precision. The shortcomings of optical character recognition (OCR) became apparent early on, as the technology struggled with degraded paper quality, unusual typefaces, and complex page layouts—common traits of newspapers more than a century old. Consequently, a vast amount of valuable information remains effectively trapped, inaccessible to both internal teams and outside researchers who cannot locate materials through keyword searches alone.

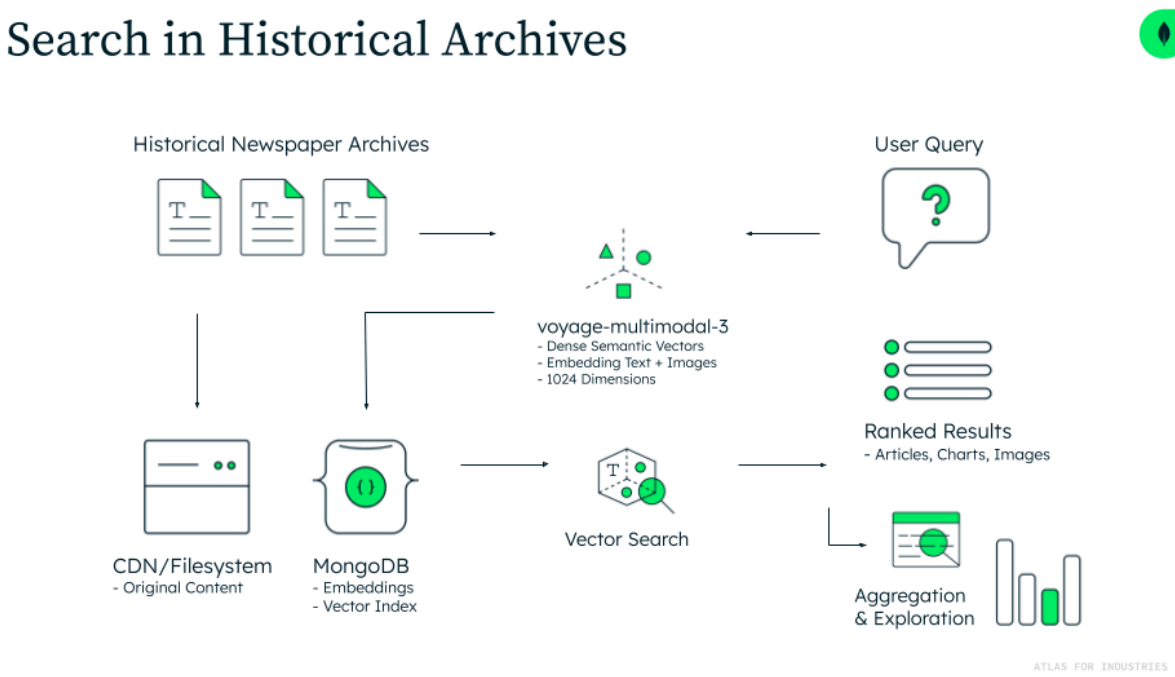
The challenge extends beyond text. Historical publications contain rich visual content—economic charts, statistical graphics, infographics, and data visualizations—that resist traditional search entirely. A researcher seeking historical commodity price trends or demographic statistics must manually page through thousands of scanned editions, as OCR cannot interpret the meaning embedded within bar charts, line graphs, or tabular data presented as images. These visual artifacts, often the most valuable elements for analysis, remain invisible to search systems.

Similar challenges have appeared in case studies from North America, Europe, Asia, and Oceania, revealing that while digitization preserves physical materials, searchability and metadata consistency remain persistent barriers. One U.S.-based museum achieved high OCR accuracy and perfect file-naming precision using flatbed scanners, specialized cradles, and deep QA workflows. Nevertheless, the fundamental issue persists: OCR output remains brittle and inconsistent across decades-spanning collections. Museums and publishers rely heavily on manual metadata enrichment and keyword search that rarely delivers the semantic richness needed for advanced research.

The Multimodal AI Breakthrough

The true shift is now emerging through artificial intelligence. Multimodal models such as Voyage-Multimodal-3 are redefining how publishers can unlock their archives. Rather than depending on OCR alone, these advanced models interpret both text and imagery directly from scans, mapping entire pages into dense semantic vectors. This approach moves beyond simple keyword search, enabling users to query vast archives by meaning, context, or visual concept. Critically, Voyage-Multimodal-3 understands the semantic content of charts, graphs, and statistical visualizations, allowing queries like "inflation trends in the 1970s" to surface relevant economic charts even when no explanatory text accompanies them.

Imagine querying an archive not only to find every front-page report on inflation in the 1970s, but also to compare how the subject was handled in news articles versus opinion pieces, and how the use of economic charts changed from one year to the next. This capability opens the door to truly statistical exploration: researchers can measure how coverage of major events or emerging issues shifted in prevalence, tone, and detail across different sections and eras. It's now possible to analyze, for example, how the narrative around nuclear energy moved from political debate in the 1960s to scientific consensus decades later, and to pinpoint whether these changes appeared first in editorial columns or investigative features.



Scaling with MongoDB Atlas

With MongoDB Atlas and state-of-the-art multimodal AI, user queries can go beyond simple keyword searches—they can encapsulate sophisticated, research-driven questions that require analyzing the statistical distribution and contextual appearance of topics throughout entire archives. For example, a user can explore not just when and where a subject like “renewable energy” was mentioned, but also uncover how frequently it appeared across different newspaper sections (such as front page, editorials, or opinion pages), which editors or writers most often discussed it, and how its presentation evolved in both textual and visual formats.

This level of analysis is possible because semantic search, enabled by vector embeddings, maps every document, page, and even individual image into a shared high-dimensional space. The workflow begins when a user formulates a query—such as measuring the rise and fall of “public transport debates” over several decades, segmented by section, region, or author. The system processes this query into its own semantic vector and compares it to the vectors of all archived artifacts. It then efficiently retrieves not only closely matching articles and charts but also aggregates results to enable statistical analysis: plotting topic

frequency by year, mapping sentiment shifts, or tracking editorial engagement across the publication.

While the semantic search and vector database infrastructure—core strengths of MongoDB—do the heavy lifting for relevance and speed, building a system for statistical or distributional research typically involves light pre- and post-processing steps. These “glue” components might extract and normalize metadata, segment results by section or author, or compute frequency histograms and time series charts. Thanks to MongoDB’s unified document model and aggregation pipelines, this surrounding logic is straightforward to implement, allowing publishers and developers to create powerful research tools that feel seamless and intuitive for end-users. In the end, MongoDB ensures that even the most complex, data-driven search and analysis workflows are not only feasible, but genuinely easy to build on top of robust, scalable infrastructure.

Roadmap for IT Leaders

Decision-makers seeking to modernize their historical archives should start by identifying high-potential pilot collections in the range of 10,000 to 20,000 pages. These collections should include a diverse variety of content types—such as articles, advertisements, statistical charts, infographics, and potentially even video segments if available—to fully validate the ability of multimodal embedding models and vector search technologies to accurately surface both textual and visual content through semantic queries. Achieving over 90% retrieval recall across varied content types, reducing labor costs significantly, accelerating research workflows, and improving user engagement with the archive—these are some of the success metrics that can be measured.

Modernizing archives with multimodal AI and vector search is not simply a project of digitization or incremental improvement. It is a strategic transformation that turns static, hard-to-navigate collections into dynamic, searchable knowledge systems. This next generation of archive management unlocks profound new value for publishers, researchers, and commercial stakeholders, supporting improved operational efficiency, enhanced cultural preservation, and expanded revenue opportunities through API licensing and the monetization of visual and long-tail data assets.

MongoDB Atlas provides the robust foundation and agile platform to realize this vision, combining cloud automation, flexible document schemas, and

comprehensive search capabilities to empower organizations to bring their archival data to life while managing costs and complexity at scale.

"Voyage-Multimodal-3: A new Multimodal Embedding Model." wandb.ai, 2024.

https://wandb.ai/byyoung3/ml-news/reports/Voyage-Multimodal-3-A-new-Multimodal-Embedding-Model--VmlldzoxMDIxNDc2Mg

"voyage-multimodal-3: all-in-one embedding model for interleaved ..." blog.voyageai.com, 2024.

https://blog.voyageai.com/2024/11/12/voyage-multimodal-3/

"Innovating with MongoDB | Customer Successes, May 2025." mongodb.com, 2025.

https://www.mongodb.com/company/blog/innovating-with-mongodb-customer-successes-may-2025