Giving Archives New Life: A Semantic Search System for Radio Barcelona SER Historical Archives (1945)

Barcelona Archives Project Team

February 2026

Abstract

This report presents a comprehensive analysis of an AI-powered archival retrieval system designed to democratize access to historical radio archives from 1945 Barcelona, Spain. Developed during a five-day intensive workshop, the project addresses critical challenges in optical character recognition (OCR), archival accessibility, and historical preservation during one of the most censored periods in Spanish history. Through the integration of modern technologies—including vector databases, large language models, and semantic search—the system enables broader access to previously hidden or difficult-to-navigate archival materials. This document explores the technical implementation, methodological considerations, ethical dimensions, and future scalability of the project.

Contents

1 Project Overview

1.1 Topic and Main Objective

This project aims to give new life to historical radio archives by enabling broader and more open access to materials that have remained difficult to access and navigate. Specifically, the system focuses on Radio Barcelona SER archives from 1945, a period characterized by intense censorship during the Franco regime in Spain. The primary objective is to reinterpret content that was previously hidden or obscured by propaganda, making these historical documents searchable, analyzable, and accessible to both researchers and the general public.

The project leverages modern artificial intelligence and semantic search technologies to bridge the gap between fragile, handwritten or typewritten archival materials and contemporary users who seek to understand this critical period of Spanish history.

1.2 Problem or Guiding Question

The central problem addressed by this project is multifaceted:

- **OCR Limitations:** Historical archives are notoriously difficult and costly to process through optical character recognition (OCR). Results are often imperfect due to handwriting variations, typewriter inconsistencies, stamps, and degraded document quality.
- **Navigation Challenges:** Traditional archival systems are hard to browse, requiring users to manually search through physical or digital copies without efficient search capabilities.
- **Quality Verification Gap:** There exists no tool that allows users to actively compare OCR'd text with original documents side by side, making it difficult to verify accuracy and identify errors or censorship markers.
- **Accessibility Barriers:** Many archives are only accessible on-site with special permissions, limiting who can engage with these historically significant materials.

The guiding question emerges: *How can we leverage modern AI technologies to make historically censored and difficult-to-access archival materials broadly available while maintaining accuracy and enabling critical analysis?*

1.3 Context

The project operates within the historical and cultural context of post-Civil War Spain, specifically focusing on 1945—a year representing the peak of censorship activities during the early Franco dictatorship. Radio Barcelona SER archives from this period contain a mixture of approved broadcasts, censored content, and propaganda materials that provide critical insight into information control mechanisms of the era.

The archives consist of documents in varying states of preservation: some typed, some handwritten, many bearing censorship stamps of ambiguous meaning. These materials are housed at the National Library of Catalonia and represent a crucial but underutilized resource for understanding Catalan and Spanish history during authoritarian rule.

The project was developed during an intensive five-day workshop in Barcelona, with collaboration from archivists at the National Library of Catalonia and Universitat Autònoma de Barcelona, who provided access, operational guidance, and historical context.

2 Working with Archives

2.1 Criteria for Selection and Organization

2.1.1 Selection Criteria

The team selected a focused subset of the Radio Barcelona SER archives based on several pragmatic and analytical criteria:

- **Temporal Focus:** Documents from the single year 1945 were chosen. This year represented the peak of censorship activities, making it historically significant while remaining manageable in scope for a short-term project.
- **Chronological Traceability:** A single-year focus enabled easier chronological tracking of events, censorship patterns, and propaganda evolution.
- **Resource Constraints:** Given computational and time limitations (five-day development window), focusing on a representative subset rather than the entire archive allowed for depth over breadth.
- **Proof of Concept:** The selected materials were sufficient to demonstrate system viability and could serve as a foundation for future scaling.

2.1.2 Data Organization

The data structure follows a systematic pipeline:

1. **OCR Processing:** Each archival document was processed through OCR systems tested across different operating systems to identify optimal configurations.
2. **Markdown Transformation:** Successfully OCR'd documents were converted into structured Markdown files, preserving both textual content and document structure. Markdown files follow a naming convention based on dates (e.g., `guiradbcn_a1945m1d1.md`).
3. **Embedding Generation:** Markdown documents were processed through Hugging Face embedding models to generate high-dimensional vector representations capturing semantic meaning.
4. **Vector Storage:** Embeddings were initially stored in a Chroma database for development and testing. The system architecture supports migration to Qdrant vector database for production deployment via the web application.
5. **Retrieval Infrastructure:** The vector database enables semantic search, allowing users to find relevant documents based on meaning rather than exact keyword matches.

2.2 Critical Decisions, Limitations, and Tensions

2.2.1 Technical Challenges

Several critical decisions emerged from the constraints of working with historical archives:

- **OCR Selection:** Identifying the best-performing OCR solution required extensive testing across operating systems (Windows, Linux, macOS). Different OCR engines produced varying results depending on document characteristics (handwritten vs. typed, stamp presence, scan quality).
- **Cost Constraints:** The team operated under strict budget limitations. Processing large volumes through commercial OCR services or cloud computing platforms was prohibitive, necessitating free-tier solutions and local processing where possible.
- **Computational Limitations:** Google Colab's resource constraints prevented processing the full archive. This limitation forced strategic selection of representative documents rather than comprehensive coverage.

2.2.2 Access and Quality Limitations

Working with archival materials presented several inherent challenges:

- **Inconsistent Editing:** Documents varied significantly in format—some typed, some handwritten, some containing mixed formats. This inconsistency complicated automated processing.
- **Censorship Ambiguity:** Stamps and markings appeared throughout documents, but their meaning was not always clear. It remained uncertain which stamps indicated censorship, approval, or administrative processing.
- **Download Restrictions:** Archives could not be downloaded in bulk. Each file required individual access, significantly slowing the data collection process.
- **Scan Quality Variation:** Document scan quality ranged from excellent to barely legible, directly impacting OCR accuracy and requiring manual verification for critical passages.

2.2.3 Ethical Considerations

Several ethical tensions emerged when digitizing and interpreting censored materials:

- **Accuracy and Interpretation:** OCR errors could potentially misrepresent historical content, especially when dealing with censored or altered documents. The team recognized the responsibility to maintain accuracy while acknowledging system limitations.
- **Censorship Detection:** The AI system cannot reliably identify which portions of documents were censored, altered, or removed. The absence of explicit censorship markers makes it impossible to determine with certainty what content was suppressed and how it should be historically contextualized.

- **Accessibility vs. Preservation:** While increasing access is valuable, the project must balance democratization with respect for archival integrity and historical context. Making materials broadly available requires careful framing to prevent misinterpretation.
- **Representation of History:** The system presents historical materials that reflect propaganda and censorship. There is an ethical responsibility to help users understand they are viewing controlled information from an authoritarian regime, not objective historical records.

2.2.4 Methodological Tensions

The project navigated several methodological tensions:

- **Breadth vs. Depth:** Limited resources forced a choice between processing many documents with less verification or processing fewer documents with greater accuracy and context. The team chose depth, focusing on a manageable subset.
- **Automation vs. Manual Curation:** While automation enabled scalability, manual verification remained necessary for quality assurance, creating tension between efficiency and accuracy.
- **Technological Determinism:** There was risk of allowing available tools to shape the research questions rather than selecting tools based on research needs. The team worked to remain problem-oriented rather than tool-oriented.

3 Process Development

3.1 Main Stages of Development

The project followed a structured yet adaptive development process across several distinct phases:

3.1.1 Stage 1: Online Preparation and System Design

The initial phase occurred remotely before the Barcelona workshop:

- **Conceptual Framework:** Team members researched archival challenges, semantic search technologies, and RAG (Retrieval-Augmented Generation) architectures.
- **System Architecture Planning:** High-level architecture was designed, including component relationships (frontend, backend, preprocessing pipeline, vector database, LLM integration).
- **Archive Understanding:** The team studied the Radio Barcelona SER archives remotely, understanding document types, historical context, and potential challenges.
- **Technology Selection:** Initial decisions were made regarding programming languages (Python for backend/processing, Vue.js for frontend), database options, and AI models.

3.1.2 Stage 2: Stakeholder Engagement in Barcelona

Upon arrival in Barcelona, direct engagement with archivists reshaped the project:

- **Archive Access:** Physical access to the National Library of Catalonia and archival materials provided concrete understanding of document conditions, formats, and access constraints.
- **Information Retrieval Challenges:** Meetings with archivists from the National Library of Catalonia and Universitat Autònoma de Barcelona identified specific pain points: difficulty browsing materials, inability to verify OCR accuracy, lack of semantic search capabilities.
- **Requirement Refinement:** Initial assumptions were validated or adjusted based on stakeholder feedback. The side-by-side comparison feature emerged as a critical need during these discussions.
- **Historical Contextualization:** Archivists provided crucial context about censorship patterns, document authenticity markers, and historical significance of 1945.

3.1.3 Stage 3: Parallel Development

Development proceeded along multiple parallel tracks:

OCR and Preprocessing Track:

- Testing multiple OCR solutions across operating systems
- Developing document cleaning and normalization pipelines
- Creating Markdown conversion utilities
- Building embedding generation workflows
- Establishing vector database population procedures

Web Application Track:

- **Frontend Development:** Building Vue.js-based user interface with Tailwind CSS, implementing chat interface, archive browsing, source viewing with side-by-side comparison
- **Backend Development:** Creating FastAPI-based REST API, implementing authentication and admin routes, building RAG retrieval logic, integrating LLM (Gemini 2.5 Flash)

Database and Infrastructure Track:

- Setting up Chroma vector database for development
- Implementing Qdrant integration for production scaling
- Containerizing services with Docker
- Creating Docker Compose orchestration

3.1.4 Stage 4: Integration and Testing

As components matured, integration became the focus:

- **Pipeline Integration:** Connecting preprocessing pipeline to web application was complex, requiring careful attention to data formats, API contracts, and error handling.
- **End-to-End Testing:** Testing complete user workflows from query submission through document retrieval and presentation.
- **Agent Behavior Testing:** Extensive testing of the LangChain-based agent to ensure appropriate responses across different query types.

3.1.5 Stage 5: Prompt Engineering and Refinement

The final stage focused on optimizing AI behavior:

- **Hallucination Reduction:** A dedicated subgroup worked intensively on prompt engineering to minimize AI hallucinations when answering archival questions.
- **API Key Management:** Frequent testing required generating multiple free-tier Gemini API keys, a time-consuming but necessary process.
- **Context Optimization:** Tuning retrieval parameters (number of documents, similarity thresholds) to balance relevance with response quality.
- **User Experience Refinement:** Final adjustments to conversational flow, response formatting, and error messaging.

3.2 Challenges Encountered and Solutions

3.2.1 Challenge 1: OCR Accuracy and Platform Dependencies

Problem: Different OCR engines produced varying results depending on the operating system and document characteristics. No single solution consistently worked well across all document types.

Solution: The team conducted systematic testing across Windows, Linux, and macOS platforms, evaluating multiple OCR engines (including open-source and commercial options). The final selection balanced accuracy, cost, and ease of integration. Document-specific preprocessing (noise reduction, contrast adjustment) was applied before OCR to improve results.

3.2.2 Challenge 2: Building RAG Architecture Under Time Pressure

Problem: Several team members were working with RAG (Retrieval-Augmented Generation) architecture for the first time. Implementing a large-scale system within five days was ambitious.

Solution: The team adopted a "must-have vs. nice-to-have" prioritization strategy. Core functionality (semantic search, basic chat, document retrieval) was implemented first. Advanced features (censorship detection, advanced filtering) were documented as future work. Leveraging existing frameworks (LangChain, FastAPI, Vue.js) rather than building from scratch accelerated development.

3.2.3 Challenge 3: Computational Resource Constraints

Problem: Google Colab limitations prevented processing the entire archive. Local machines lacked sufficient GPU memory for large-scale embedding generation.

Solution: The team strategically selected a representative subset (1945 documents) that was both historically significant and computationally manageable. Architecture was designed for future scalability, with clear documentation of requirements for full archive processing (estimated 18GB GPU memory needed).

3.2.4 Challenge 4: Integrating Preprocessing Pipeline with Web Application

Problem: The preprocessing pipeline (data collection, OCR, embedding) and web application (query handling, retrieval, presentation) were developed somewhat independently, creating integration challenges.

Solution: The team established clear data contracts and API specifications early in the process. Regular integration checkpoints ensured components remained compatible. Docker containerization isolated components while enabling straightforward deployment.

3.2.5 Challenge 5: Gemini API Key Management

Problem: Frequent testing and prompt tuning quickly exhausted free-tier API quotas, requiring generation of multiple new API keys, which was tedious and time-consuming.

Solution: The team implemented API key rotation in configuration, allowing quick switching between keys. Usage was carefully monitored to maximize efficiency. The architecture was designed to support multiple LLM providers (OpenAI, Claude, local models), reducing dependency on any single provider.

3.2.6 Challenge 6: Hallucination Control

Problem: The LLM occasionally generated plausible-sounding but inaccurate information about archival content, particularly when asked about topics not directly covered in retrieved documents.

Solution: A dedicated prompt engineering subgroup systematically refined system prompts to emphasize reliance on retrieved documents. The agent was instructed to explicitly acknowledge uncertainty when relevant documents weren't found. Response templates encouraged citation of specific source documents.

3.3 Key Learnings from the Process

3.3.1 Technical Learnings

- **RAG Architecture Viability:** RAG-based systems can be successfully implemented in short timeframes when properly scoped, providing significant value for domain-specific question-answering tasks.

- **Vector Database Performance:** Modern vector databases (Chroma, Qdrant) enable fast semantic search even with limited optimization, making them practical for archival applications.

- **Embedding Model Selection:** Choice of embedding model significantly impacts retrieval quality. Models must be evaluated specifically for the target language (Spanish/Catalan) and domain (historical documents).
- **OCR as Critical Bottleneck:** OCR quality remains the fundamental constraint in historical archive digitization. Improvements in OCR technology would have multiplicative effects on downstream applications.

3.3.2 Process Learnings

- **Stakeholder Engagement Value:** Direct engagement with archivists transformed the project from a technical exercise into a practical tool addressing real needs. Early and continuous stakeholder involvement is essential.
- **Parallel Development Effectiveness:** Dividing the team into specialized subgroups (OCR/preprocessing, web application, prompt engineering) enabled efficient progress while maintaining integration touchpoints.
- **Problem-Oriented Over Tool-Oriented:** Maintaining focus on actual archival challenges rather than showcasing technologies led to more practical outcomes. The question "What problem are we solving?" guided critical decisions.
- **Iterative Refinement Necessity:** Initial implementations were rarely optimal. Building in time for iterative testing and refinement proved essential, particularly for prompt engineering.

3.3.3 Domain Learnings

- **Historical Context Matters:** Understanding the historical context of censorship during Franco's regime was essential for making appropriate design decisions and interpreting results.
- **Archival Complexity:** Archives are far more complex than digital-native documents. Inconsistencies, ambiguities, and physical degradation require flexible, robust processing pipelines.
- **Access Democratization Impact:** Even imperfect tools can significantly democratize access to previously inaccessible materials, creating value for diverse user communities.

3.3.4 Collaborative Learnings

- **Interdisciplinary Collaboration:** Combining technical expertise with historical/archival knowledge produced better outcomes than either discipline alone could achieve.
- **Communication Overhead:** Short timelines required efficient communication. Daily standups, clear task assignments, and shared documentation prevented duplication and confusion.

- **Expertise Distribution:** Team members brought varying levels of experience with AI, web development, and archival work. Pairing experienced members with learners accelerated knowledge transfer.

4 Technological Dimension

4.1 Technologies Used

The project integrates a diverse technology stack spanning multiple domains:

4.1.1 Programming Languages

- **Python 3.x:** Primary language for backend API, preprocessing pipeline, and data processing
- **JavaScript (ES6+):** Frontend development with Vue.js framework
- **SQL:** Vector database queries and metadata storage

4.1.2 Frontend Technologies

- **Vue.js 3:** Progressive JavaScript framework for building the user interface
- **Vite:** Fast build tool and development server
- **Tailwind CSS:** Utility-first CSS framework for styling
- **PostCSS:** CSS processing and optimization

4.1.3 Backend Technologies

- **FastAPI:** Modern Python web framework for building REST APIs
- **Uvicorn:** ASGI server for serving the FastAPI application
- **Pydantic:** Data validation and settings management

4.1.4 AI and Machine Learning

- **Large Language Models:**
  - Primary: Google Gemini 2.5 Flash for chat and question-answering
  - Architecture supports: OpenAI GPT models, Anthropic Claude, local models (planned)
- **Embedding Models:**
  - Multiple Hugging Face models tested and deployed
  - One model for initial testing and experimentation
  - Different model selected for final production pipeline

  - Models specifically chosen for Spanish/Catalan language support

- **LangChain:** Framework for orchestrating LLM workflows, agent behavior, and RAG implementation
- **CLIP Handler:** Legacy component from previous project, retained for text embedding handling (originally designed for image-based archive search)

4.1.5 Vector Databases

- **Chroma:** Used for development, testing, and initial prototyping. Offers simplicity and ease of setup for local development.
- **Qdrant:** Designated for production deployment via web application. Provides better scalability, performance, and features for production environments.

4.1.6 OCR and Document Processing

- Multiple OCR engines tested (specific names determined through cross-platform evaluation)
- Image preprocessing libraries (OpenCV, PIL/Pillow) for document enhancement
- Text cleaning and normalization utilities
- Markdown generation tools for structured output

4.1.7 Infrastructure and Deployment

- **Docker:** Containerization of all services (frontend, backend, preprocessing, databases)
- **Docker Compose:** Multi-container orchestration for local development and deployment
- **Nginx:** Web server for serving frontend static files and reverse proxying

4.1.8 Development Tools and Platforms

- **Google Colab:** Used for initial experimentation and testing (with noted limitations)
- **Jupyter Notebooks:** Data exploration and pipeline development
- **Git:** Version control (implied by modern development practices)

4.2 Technology Process

4.2.1 Data Ingestion and Processing Pipeline

The technology process follows a multi-stage pipeline:

Stage 1: Document Acquisition

1. Individual document access from archival sources (no bulk download available)
2. Manual or semi-automated downloading of scanned PDFs or images
3. Organization by date and document type

Stage 2: OCR Processing

1. Image preprocessing (noise reduction, contrast enhancement, deskewing)
2. OCR engine application across different platforms
3. Text extraction with confidence scoring
4. Error detection and flagging for manual review

Stage 3: Text Normalization and Structuring

1. Text cleaning (removing OCR artifacts, correcting encoding issues)
2. Markdown conversion with structural preservation
3. Metadata extraction (dates, document types, censorship markers when identifiable)
4. File naming according to convention (e.g., `guiradbcn_a1945m1d1.md`)

Stage 4: Embedding Generation

1. Loading processed Markdown documents
2. Text chunking strategies (balancing context window size with semantic coherence)
3. Embedding generation using selected Hugging Face models
4. High-dimensional vector creation capturing semantic meaning

Stage 5: Vector Database Population

1. Initial storage in Chroma database for development
2. Migration capability to Qdrant for production
3. Metadata indexing alongside embeddings
4. Collection organization and index optimization

4.2.2 Query Processing and Retrieval Workflow

User Query Flow

1. User submits query through Vue.js frontend chat interface
2. Frontend sends HTTP request to FastAPI backend
3. Backend receives query and routes to appropriate handler

Agent Decision Making

1. LangChain agent analyzes query type and intent
2. **Casual Conversation:** For greetings or casual queries, agent responds directly with minimal token usage
3. **Archive Queries:** For substantive historical questions, agent triggers retrieval process
4. **Follow-up Questions:** Agent maintains conversation context and can reuse previously retrieved documents from chat history

Semantic Retrieval Process

1. Query text is embedded using the same model used for document embeddings
2. Vector similarity search performed against database (cosine similarity or similar metric)
3. Top-k most relevant documents retrieved based on similarity scores
4. Relevance thresholds applied to filter low-quality matches

RAG-Based Response Generation

1. Retrieved documents injected into LLM context window
2. System prompt instructs LLM to:
   - Ground responses in retrieved documents
   - Acknowledge uncertainty when information is unavailable
   - Cite sources when possible
   - Avoid hallucination by staying within document scope
3. Gemini 2.5 Flash generates natural language response
4. Response includes conversational answer and source document references

Frontend Presentation

1. Backend returns JSON response with answer and source metadata
2. Frontend displays answer in chat interface
3. Source documents presented with links to full view
4. Side-by-side comparison available (OCR text vs. original scan) through `SourceView.vue` component

4.2.3 Containerized Deployment Architecture

The Docker-based architecture ensures reproducibility and portability:

- **Frontend Container:** Nginx serving Vue.js static files, reverse proxying API requests
- **Backend Container:** FastAPI application with Python dependencies, LangChain integration
- **Vector Database Container:** Qdrant (or Chroma) with persistent volume storage
- **Preprocessing Container:** Isolated environment for OCR and embedding generation (can be run independently)
- **Docker Compose:** Orchestrates all services, manages networking, environment variables, and volume mounts

4.3 Role of Technology in Meaning-Making

Technology plays multiple critical roles in transforming raw archival materials into accessible knowledge:

4.3.1 Democratization of Access

- **Breaking Physical Barriers:** Digital access eliminates the need for on-site visits and special permissions, enabling global access to localized historical materials.
- **Search Beyond Keywords:** Semantic search allows users to find relevant documents based on conceptual meaning rather than exact keyword matches, which is essential for historical research where terminology may vary.
- **Conversational Interface:** Natural language chat interface lowers technical barriers, allowing historians, students, and general public to engage with archives without specialized search syntax.

4.3.2 Analytical Capabilities

- **Pattern Recognition:** Vector embeddings enable identification of thematic patterns across documents that might not be apparent through traditional reading.
- **Comparative Analysis:** Users can quickly compare multiple documents, identifying similarities and differences in censorship approaches, propaganda narratives, or content evolution.
- **Context Synthesis:** RAG architecture synthesizes information across multiple documents, providing comprehensive answers that would require hours of manual cross-referencing.

4.3.3 Preservation and Longevity

- **Digital Preservation:** OCR and digital storage protect content from physical degradation of original documents.
- **Format Flexibility:** Markdown and structured data formats ensure long-term accessibility independent of proprietary software.
- **Redundancy:** Digital copies can be easily duplicated and distributed, reducing risk of irreversible loss.

4.3.4 Critical Engagement

- **Side-by-Side Verification:** The ability to compare OCR'd text with original scans enables critical assessment of accuracy and identification of potential misreadings that could alter historical interpretation.
- **Transparency of Uncertainty:** System design encourages acknowledgment of limitations, ambiguities, and gaps in knowledge, promoting critical historical thinking rather than false certainty.
- **Revealing Hidden Content:** By making censored materials searchable and analyzable, technology enables research into what was suppressed, how control mechanisms operated, and what propaganda narratives were promoted.

4.4 Level of Dependence or Autonomy

The project exhibits both dependencies and autonomous capabilities:

4.4.1 Current Dependencies

- **Commercial LLM API:** Primary dependency on Google Gemini API for response generation. While free-tier usage is possible, scaling requires paid access. API changes or service discontinuation would impact functionality.
- **Embedding Model Hosting:** Hugging Face models require either cloud inference or local GPU resources. Limited local hardware creates dependency on external compute.
- **OCR Quality:** Fundamental dependence on OCR technology accuracy. Poor OCR results cascade through the entire pipeline, limiting system effectiveness.
- **Initial Manual Labor:** Document selection, preprocessing, and quality verification require human judgment and effort. Fully automated processing remains unachievable for complex historical materials.

4.4.2 Autonomy and Flexibility

- **LLM Provider Flexibility:** Architecture supports multiple LLM providers (Gemini, OpenAI, Claude, local models). This modularity reduces vendor lock-in.

- **Local Deployment Capability:** Entire system can theoretically run locally with sufficient hardware (estimated 18GB GPU memory), eliminating cloud dependencies.
- **Open-Source Foundation:** Use of open-source embedding models (Hugging Face), frameworks (LangChain, FastAPI, Vue.js), and databases (Chroma, Qdrant) ensures community support and modification capability.
- **Data Ownership:** All processed data (Markdown files, embeddings, metadata) is locally stored and portable, preventing vendor lock-in.

4.4.3 Technical Debt and Future Autonomy

Stakeholders encouraged migration to fully local models, which remains unimplemented due to time constraints:

- **Local LLM Deployment:** Transitioning to local models (e.g., LLaMA, Mistral) would eliminate API dependencies and costs while enabling customization for Spanish/Catalan historical text.
- **Custom Fine-Tuning:** Local models could be fine-tuned on historical Spanish documents and censorship patterns, potentially improving performance for domain-specific tasks.
- **Privacy and Control:** Local deployment ensures complete data privacy and institutional control, critical for sensitive historical materials.

4.5 Technical Constraints, Biases, and Limitations

4.5.1 AI Model Limitations

- **Censorship Detection Failure:** The most significant limitation is the AI's inability to reliably detect which portions of documents were censored, altered, or removed. Without explicit censorship markers, the system cannot distinguish between original content and propaganda modifications. This affects:
  - Historical accuracy and interpretation
  - Understanding of censorship mechanisms
  - Reconstruction of suppressed narratives
- **Hallucination Risk:** Despite prompt engineering efforts, LLMs can generate plausible but inaccurate information, particularly when queries fall outside the scope of retrieved documents.
- **Language Bias:** Most LLMs are optimized for English. Performance on Spanish and Catalan historical text (with period-specific vocabulary and syntax) may be suboptimal.
- **Historical Context Understanding:** LLMs lack deep understanding of 1945 Spanish political context, potentially leading to anachronistic interpretations or missed nuances.

4.5.2 OCR Constraints

- **Handwriting Recognition:** Handwritten documents yield significantly lower OCR accuracy than typed materials, sometimes approaching illegibility.
- **Mixed-Format Documents:** Documents combining typed text, handwritten annotations, and stamps create complex OCR challenges that current technology handles imperfectly.
- **Degradation Effects:** Physical degradation (fading, tears, stains) directly reduces OCR success rates.
- **Stamp Interference:** Censorship stamps and administrative markings often overlap text, obscuring content and confusing OCR systems.

4.5.3 Computational Constraints

- **Memory Limitations:** High-dimensional embeddings and large context windows require substantial GPU memory (18GB estimated for full archive), limiting deployment options.
- **Processing Time:** Embedding generation for large archives is time-intensive, creating bottlenecks for scaling.
- **Cost Barriers:** Commercial LLM APIs incur per-token costs that scale with usage, potentially limiting accessibility for under-funded institutions.

4.5.4 Retrieval Limitations

- **Embedding Quality Dependency:** Retrieval quality is bounded by embedding model performance. Poor embeddings lead to irrelevant retrievals regardless of retrieval algorithm sophistication.
- **Chunk Size Trade-offs:** Smaller chunks improve specificity but lose context; larger chunks preserve context but reduce precision. Optimal chunking remains document-dependent.
- **Similarity Metric Limitations:** Cosine similarity and similar metrics capture semantic similarity but may miss relevant documents with different linguistic expression of similar concepts.

4.5.5 Systematic Biases

- **Training Data Bias:** Embedding and LLM models trained primarily on contemporary, English-dominated corpora may underperform on historical Spanish/Catalan censored materials.
- **Selection Bias:** Focusing on 1945 provides depth but may not represent patterns from other years. Generalizations must be made cautiously.

- **Preservation Bias:** Surviving archival materials may not represent all content produced; censorship and selective preservation create inherent bias in available sources.
- **Accessibility Bias:** Digital access benefits those with internet connectivity and technological literacy, potentially excluding communities with historical connections to these materials.

4.5.6 Methodological Limitations

- **Verification Challenge:** Assessing system accuracy requires domain expertise. Non-expert users may trust incorrect responses without ability to verify.
- **Context Dependency:** System performance varies significantly based on query formulation, retrieved document relevance, and prompt engineering quality.
- **Evaluation Difficulty:** Measuring success in historical interpretation is inherently subjective, making quantitative evaluation challenging.

5 Methodologies and Tools

5.1 Methodologies Applied

The project integrated multiple methodological approaches rather than adhering to a single rigid framework:

5.1.1 Problem-Oriented Approach

The overarching methodology was **problem-oriented** rather than methodology-driven:

- **Starting Point:** Real archival challenges (access, browsing, verification) identified through stakeholder engagement drove all technical decisions.
- **Constraint-Driven:** Cost and time limitations around OCR strongly influenced architectural choices and scope definition.
- **Pragmatic Solutions:** Technology selection prioritized practical solutions over theoretical elegance or showcasing cutting-edge tools.
- **Adaptive Decision-Making:** When initial approaches failed (e.g., Colab computational limits), the team pivoted quickly to viable alternatives rather than adhering to predetermined plans.

5.1.2 Hybrid Waterfall-Agile Process

The development process combined elements of waterfall and agile methodologies:

Waterfall Elements:

- **Upfront Architecture Design:** System architecture was largely designed before implementation began during the online preparation phase.
- **Component Specification:** Clear interfaces and data contracts between components (frontend, backend, preprocessing) were defined early.
- **Linear Pipeline Stages:** Data processing followed a sequential pipeline (acquisition → OCR → embedding → storage → retrieval).

Agile Elements:

- **Iterative Development:** Implementation remained flexible and adaptive, especially after stakeholder engagement in Barcelona.
- **Daily Standups:** Team coordination through brief daily meetings to sync progress and address blockers.
- **Prioritized Backlog:** Features classified as "must-have" vs. "nice-to-have," with continuous re-prioritization based on progress and constraints.
- **Rapid Prototyping:** Quick prototypes tested assumptions before full implementation.

5.1.3 Exploratory and Experimental Methods

Significant exploratory work characterized early stages:

- **OCR Experimentation:** Systematic testing of multiple OCR solutions across different platforms to identify best performers.
- **Embedding Model Evaluation:** Trying multiple Hugging Face models with test queries to assess retrieval quality before committing to production model.
- **Prompt Engineering Iteration:** Continuous experimentation with different prompts, instructions, and context structures to minimize hallucinations and improve response quality.
- **Parameter Tuning:** Exploring retrieval parameters (number of documents retrieved, similarity thresholds, chunking strategies) through trial and error.

5.1.4 Collaborative and Interdisciplinary Methodology

The project inherently required collaboration across disciplines:

- **Team-Based Development:** Work divided among specialized subgroups (OCR/preprocessing, web application, prompt engineering) with regular integration points.
- **Stakeholder Co-Design:** Archivists from the National Library of Catalonia and Universitat Autònoma de Barcelona actively shaped requirements and validated approaches.

- **Knowledge Sharing:** Experienced team members mentored others new to RAG architecture, fostering collective capability growth.
- **Interdisciplinary Synthesis:** Combining technical AI/software engineering expertise with historical and archival domain knowledge to create contextually appropriate solutions.

5.1.5 Analytical Methodology

Research and analysis informed multiple project dimensions:

- **Historical Analysis:** Studying 1945 Spanish censorship patterns to understand document characteristics and research needs.
- **Comparative Technology Analysis:** Evaluating multiple OCR engines, embedding models, vector databases, and LLM providers through structured comparison.
- **User Needs Analysis:** Identifying pain points through stakeholder interviews and observing archival researcher workflows.
- **Performance Evaluation:** Testing retrieval quality, response accuracy, and system performance under various conditions.

5.2 Relationship Between Methodology, Objectives, and Outcomes

5.2.1 Alignment: Problem-Oriented Approach ¡-¿ Practical Outcomes

The problem-oriented methodology directly enabled practical outcomes:

- **Objective:** Enable broader access to difficult-to-navigate archives
- **Methodology:** Focus on real archival challenges identified by stakeholders
- **Outcome:** System features (semantic search, side-by-side comparison, conversational interface) directly address identified pain points rather than generic capabilities

5.2.2 Alignment: Hybrid Process ¡-¿ Time Constraints

The waterfall-agile hybrid suited the five-day timeline:

- **Objective:** Deliver functional prototype within workshop timeframe
- **Methodology:** Upfront architecture (waterfall) provided clear direction; iterative implementation (agile) allowed adjustment to discoveries and blockers
- **Outcome:** Working system delivered on time despite challenges, with clear documentation of future work for features not completed

5.2.3 Alignment: Exploratory Methods ¡-¿ Technical Uncertainty

Experimentation addressed areas of high uncertainty:

- **Objective:** Select optimal technologies for Spanish/Catalan historical documents
- **Methodology:** Systematic experimentation with OCR engines, embedding models, and prompt strategies
- **Outcome:** Evidence-based technology choices rather than assumptions, improving system performance and reliability

5.2.4 Alignment: Collaborative Methods ¡-¿ Interdisciplinary Nature

Team collaboration matched the interdisciplinary challenge:

- **Objective:** Create historically informed, technically sound archival tool
- **Methodology:** Sustained engagement with archivists; knowledge sharing within technical team
- **Outcome:** System respects archival principles and historical context while leveraging modern AI capabilities; team capacity increased through knowledge transfer

5.2.5 Tensions and Trade-offs

Some methodological choices created trade-offs:

- **Breadth vs. Depth:** Problem-oriented focus on 1945 provided depth and proof of concept but limited generalization to other years/archives. This trade-off was conscious and appropriate given constraints.
- **Speed vs. Perfection:** Five-day timeline necessitated "good enough" solutions rather than optimal ones. For example, prompt engineering reduced but didn't eliminate hallucinations; OCR accuracy improved but remained imperfect.
- **Autonomy vs. Expediency:** Using commercial API (Gemini) enabled rapid development but created dependency. Methodology prioritized immediate functionality over long-term autonomy, with documented path to local models as future work.

5.2.6 Lessons on Methodology-Objective Fit

Key insights on methodology effectiveness:

- **Context Matters:** The hybrid waterfall-agile approach worked well for a short-term, bounded project with clear objectives. Longer projects might benefit from more iterative architecture evolution.
- **Stakeholder Engagement is Essential:** Without archivists' input, the project would have built generic capabilities missing actual user needs. Co-design methodology proved invaluable.

- **Experimentation Requires Time:** Exploratory phases (OCR testing, model evaluation) were time-consuming but essential. Allocating sufficient time for experimentation improved final outcomes.
- **Documentation as Methodology:** Documenting technical debt and future work proved crucial for project sustainability. Methodology should explicitly include knowledge capture, not just implementation.

6 Proposed Use and Real-World Application

6.1 Projection Beyond the Classroom

While developed as an intensive workshop project, the system is designed with real-world deployment in mind:

6.1.1 Immediate Viability

The system is operational and ready for limited deployment:

- **Functional Prototype:** All core features (semantic search, chat interface, document retrieval, side-by-side comparison) are implemented and tested.
- **Containerized Deployment:** Docker-based architecture enables straightforward deployment on institutional servers or cloud platforms.
- **Representative Dataset:** 1945 Radio Barcelona SER archives provide sufficient content for meaningful research and public engagement.

6.1.2 Scalability Path

Clear pathways exist for expanding beyond the initial scope:

- **Temporal Expansion:** Methodology can be applied to additional years (1940s-1950s) using the same preprocessing pipeline.
- **Archival Expansion:** System architecture accommodates other archival collections beyond Radio Barcelona (newspapers, government documents, personal correspondence).
- **Geographic Expansion:** Approach could be adapted for archives from other regions of Spain or other countries with censored historical materials.

6.2 Potential Contexts of Application

6.2.1 Educational Contexts

Universities and Research Institutions

- **Graduate Research:** History, political science, and communication studies students conducting research on Franco-era Spain, censorship mechanisms, or propaganda analysis.

- **Digital Humanities Courses:** Demonstrating intersection of computational methods and historical inquiry; teaching critical evaluation of AI-assisted research.
- **Archival Studies Programs:** Case study for modern archival practices, digitization challenges, and ethical considerations in archive accessibility.

Secondary Education

- **Spanish History Curriculum:** Engaging students with primary sources from Civil War and dictatorship periods through accessible digital interface.
- **Critical Media Literacy:** Analyzing propaganda techniques and censorship patterns; understanding historical information control.
- **Technology Education:** Introducing students to AI applications in humanities; discussing capabilities and limitations of automated analysis.

6.2.2 Cultural Contexts

Libraries and Archives

- **National Library of Catalonia:** Primary stakeholder; system could augment existing digital archive access, providing semantic search capabilities currently unavailable.
- **Municipal Archives:** Local historical societies and municipal archives across Catalonia and Spain with similar materials.
- **Specialized Collections:** Radio archives, newspaper collections, and other periodical materials from mid-20th century Spain.

Museums and Cultural Institutions

- **Historical Memory Museums:** Institutions focused on Spanish Civil War and Franco dictatorship (e.g., MUHBA - Barcelona History Museum) could integrate system for public research stations.
- **Interactive Exhibits:** Public-facing terminals allowing museum visitors to explore archival materials through conversational interface.
- **Digital Exhibitions:** Online exhibitions incorporating semantic search for curated archival collections.

6.2.3 Professional Contexts

Academic Research

- **Historians:** Researchers studying 1940s Spain, censorship history, media control, or propaganda can conduct more efficient archival research.
- **Linguists:** Analyzing language use, propaganda rhetoric, and censored discourse patterns.

- **Political Scientists:** Studying authoritarian information control mechanisms and their evolution.

Journalism and Media

- **Investigative Journalism:** Reporters researching historical context for contemporary stories about Spanish politics, historical memory, or media freedom.
- **Documentary Production:** Filmmakers and content creators sourcing primary materials for historical documentaries.
- **Fact-Checking:** Verifying historical claims or contextualizing contemporary debates about Spanish history.

6.2.4 Social Contexts

Public Engagement

- **Historical Memory Movements:** Organizations working on recuperating historical memory in Spain; families researching relatives' experiences during Franco regime.
- **Tourism:** Culturally engaged tourists learning about Barcelona and Catalan history through accessible digital archives.
- **Citizen Historians:** General public interested in local history, family genealogy, or understanding their community's past.

6.3 Intended Users or Audiences

6.3.1 Primary User Groups

Professional Archivists and Librarians

- **Needs:** Efficient cataloging, patron support, collection management
- **Use Cases:** Helping patrons locate relevant materials; identifying collection gaps; conducting archival research
- **Benefits:** Reduced time answering repetitive queries; improved access to collection contents

Academic Researchers

- **Needs:** Comprehensive source discovery, efficient information extraction, citation support
- **Use Cases:** Literature review; primary source analysis; comparative studies across documents
- **Benefits:** Faster research; ability to identify patterns across large document sets; access without physical archive visits

Students (University and Secondary)

- **Needs:** Accessible primary sources, guided learning, engaging interfaces
- **Use Cases:** Research projects; historical essays; critical analysis assignments
- **Benefits:** Direct engagement with primary sources; understanding historical research methods; developing critical thinking about sources

6.3.2 Secondary User Groups

Museum Professionals

- **Needs:** Content for exhibitions, educational programming, public engagement
- **Use Cases:** Curating digital exhibitions; creating educational materials; supporting public research
- **Benefits:** Rich content source; ability to create thematic explorations; engagement with museum visitors

Journalists and Media Professionals

- **Needs:** Quick research, historical context, primary source verification
- **Use Cases:** Fact-checking historical claims; sourcing documentary materials; background research
- **Benefits:** Rapid access to primary sources; ability to verify historical narratives

General Public and History Enthusiasts

- **Needs:** Accessible interfaces, engaging content, self-guided learning
- **Use Cases:** Learning local history; family research; exploring personal interests in Spanish/Catalan history
- **Benefits:** Democratized access to cultural heritage; conversational interface lowers barriers; discovering previously unknown historical details

6.4 Value or Impact in Real-World Settings

6.4.1 Access and Democratization

- **Geographic Accessibility:** Removes requirement for physical presence in Barcelona/Catalonia, enabling international researchers and diaspora communities to engage with Catalan history.
- **Economic Accessibility:** Eliminates travel costs, accommodation expenses, and potential research permissions fees that create barriers for independent researchers and under-funded institutions.

- **Temporal Accessibility:** 24/7 availability removes constraints of archive operating hours, particularly valuable for international users in different time zones.
- **Technical Accessibility:** Conversational interface is more accessible than specialized archival search systems, lowering technical barriers for non-expert users.

6.4.2 Research Acceleration

- **Discovery Efficiency:** Semantic search dramatically reduces time spent identifying relevant documents compared to manual browsing or keyword search.
- **Cross-Document Analysis:** System can synthesize information across multiple documents, enabling pattern recognition and comparative analysis that would be time-prohibitive manually.
- **Hypothesis Testing:** Researchers can quickly test preliminary hypotheses about censorship patterns, propaganda themes, or content evolution.

6.4.3 Preservation

- **Physical Preservation:** Digital access reduces handling of fragile originals, extending their physical lifespan.
- **Content Preservation:** OCR and digital storage create backup copies protecting against disaster, deterioration, or loss.
- **Interpretive Preservation:** System documentation captures archivists' knowledge about collection organization, censorship patterns, and historical context.

6.4.4 Educational Value

- **Primary Source Engagement:** Students gain direct experience with historical materials rather than relying solely on textbook summaries.
- **Critical Thinking Development:** Side-by-side comparison feature teaches students to critically evaluate OCR accuracy and source reliability.
- **Digital Literacy:** Using the system develops skills in AI-assisted research while learning to recognize AI limitations and potential biases.

6.4.5 Cultural and Social Impact

- **Historical Memory:** Facilitates recuperation of historical memory in post-Franco Spain, supporting ongoing efforts to understand and acknowledge dictatorship-era repression.
- **Transparency:** Making censored materials accessible promotes transparency about historical information control mechanisms.
- **Community Connection:** Enables Catalan diaspora and descendant communities to connect with their historical and cultural roots.

- **Counter-Narrative Development:** By revealing censored content, system supports development of historical narratives that challenge official propaganda accounts.

6.4.6 Institutional Benefits

- **Increased Visibility:** Digital accessibility raises profile of National Library of Catalonia collections, potentially attracting new users and research interest.
- **Resource Efficiency:** Reduces staff time spent on repetitive reference questions and simple queries.
- **Collection Understanding:** Semantic search capabilities help archivists themselves better understand collection contents and identify preservation or cataloging priorities.
- **Grant and Funding Opportunities:** Demonstrates institutional innovation and digital capacity, strengthening applications for digitization and technology grants.

7 Steps Toward Implementation

7.1 Requirements for Real-World Functionality

7.1.1 Technical Infrastructure Requirements

Computational Resources

- **GPU Hardware:** Estimated minimum 18GB GPU memory for processing embeddings and running queries against full archive (based on extrapolation from 1945 subset). Actual requirements depend on total archive size, which remains uncertain.
- **Server Infrastructure:** Dedicated server or cloud infrastructure capable of:
  - Running Docker containers (frontend, backend, vector database)
  - Handling concurrent user queries (scale dependent on expected traffic)
  - Storing vector database (size dependent on archive scope)
  - Serving original document scans for side-by-side comparison
- **Storage:** Sufficient storage for:
  - Original document scans (PDFs or images)
  - OCR'd Markdown files
  - Vector embeddings
  - Database indices and metadata
  - Application logs and backups
- **Network Bandwidth:** Adequate bandwidth for serving high-resolution document images and handling API traffic.

Software and Licensing

- **Open-Source Components:** All core software (Python, FastAPI, Vue.js, Docker, Nginx) uses permissive open-source licenses requiring no fees.
- **Embedding Models:** Must use open-source models from Hugging Face (already implemented) to avoid licensing costs and restrictions.
- **LLM Options:**
  - Current: Gemini API requires ongoing costs at scale (free tier insufficient for production)
  - Preferred: Local open-source models (LLaMA, Mistral) eliminate API costs but require more GPU memory
  - Alternative: OpenAI or Claude APIs as fallback, with associated costs

7.1.2 Data Requirements

Archive Expansion

- **Temporal Expansion:** Extending beyond 1945 to cover broader periods (e.g., entire 1940s decade, or 1939-1975 Franco era).
- **Collection Expansion:** Including additional archival collections beyond Radio Barcelona SER:
  - Other radio stations and broadcast archives
  - Newspaper collections from the period
  - Government censorship records
  - Personal correspondence and memoirs
- **Document Diversity:** Ensuring representation of different document types, censorship patterns, and perspectives.

Quality Assurance

- **OCR Verification:** Manual spot-checking of OCR accuracy across representative sample; correction of critical errors.
- **Metadata Enrichment:** Standardized metadata (dates, document types, authors, censorship markers) to enable filtering and advanced search.
- **Historical Annotation:** Expert annotation identifying known censorship patterns, propaganda themes, and historically significant passages.

7.1.3 Human Resources

Technical Staff

- **System Administrator:** Managing server infrastructure, database maintenance, backups, security updates.
- **Developer/Engineer:** Ongoing development (bug fixes, feature additions, performance optimization), prompt engineering refinement.
- **Data Engineer:** OCR pipeline management, embedding generation for new documents, database optimization.

Content Staff

- **Archivist/Curator:** Selecting materials for inclusion, providing historical context, quality assurance of OCR and metadata.
- **Historian:** Domain expertise for interpreting censorship patterns, validating system responses, creating educational content.
- **Metadata Specialist:** Standardizing and enriching document metadata, ensuring consistency across collections.

Support Staff

- **User Support:** Responding to user questions, troubleshooting access issues, gathering feedback.
- **Training Coordinator:** Developing user documentation, conducting workshops for researchers, creating educational materials.

7.1.4 Legal and Ethical Requirements

Copyright and Licensing

- **Archival Rights:** Confirming National Library of Catalonia has rights to digitize and distribute archive materials; some materials may have copyright restrictions despite age.
- **Privacy Considerations:** Determining if any archival materials contain personal information requiring redaction or access restrictions under GDPR or other privacy regulations.
- **Terms of Use:** Establishing clear terms of use for system access, specifying permitted uses (research, education) and restrictions (commercial use).

Ethical Frameworks

- **Accuracy Disclaimers:** Clear communication about OCR limitations, potential errors, and AI capabilities/limitations.

- **Historical Context:** Providing adequate framing so users understand materials reflect propaganda and censorship, not objective historical records.
- **Accessibility Standards:** Ensuring system complies with web accessibility standards (WCAG) for users with disabilities.

7.2 Necessary Adjustments or Future Developments

7.2.1 Technical Debt Resolution

High Priority

- **Local LLM Migration:** Transitioning from Gemini API to local open-source models:
  - Eliminates API costs and dependencies
  - Enables fine-tuning for Spanish/Catalan historical text
  - Provides institutional control and data privacy
  - Encouraged by stakeholders but not implemented due to time constraints
- **Production Database Migration:** Completing migration from Chroma (development) to Qdrant (production) with:
  - Optimized index structures
  - Performance tuning for concurrent queries
  - Backup and recovery procedures
- **PDF Embedding and Display:** Implementing direct PDF viewing within web application:
  - Currently redirects to external websites
  - Should display source PDFs directly in `SourceView.vue`
  - Improves user experience and reduces dependency on external hosting

Medium Priority

- **Censorship Detection Development:** Research and implement capabilities for:
  - Identifying censorship stamps automatically (computer vision)
  - Detecting altered or removed content patterns (NLP analysis)
  - Reconstructing historical context of censored passages
  - Visualizing censorship patterns across documents
- **Embedding Strategy Evaluation:** Comparing performance of:
  - Current approach: Markdown-based embeddings (structure + text)
  - Alternative: Pure text embeddings
  - Hybrid: Multiple embedding strategies for different query types

Determining optimal approach for different use cases.

- **Authentication and Authorization:** Implementing user accounts with:
  - Different access levels (public, researcher, administrator)
  - Usage tracking for capacity planning
  - Personalized features (saved searches, annotations)

Lower Priority

- **Advanced Filtering:** Adding filters for:
  - Date ranges
  - Document types
  - Censorship indicators
  - OCR confidence levels
- **Export Capabilities:** Enabling users to:
  - Export search results and citations
  - Download OCR'd text
  - Create research collections
- **Multilingual Interface:** Providing UI in:
  - Catalan (primary)
  - Spanish
  - English (for international researchers)

7.2.2 Quality Improvements

OCR Enhancement

- **Manual Correction Workflow:** Developing systematic process for identifying and correcting critical OCR errors.
- **Confidence Scoring:** Implementing OCR confidence scoring to flag low-confidence passages for review.
- **Handwriting Specialists:** Engaging paleography experts for particularly difficult handwritten documents.

Response Quality

- **Continued Prompt Engineering:** Ongoing refinement based on user feedback and observed failure modes.
- **Retrieval Tuning:** Optimizing retrieval parameters based on actual usage patterns.
- **Answer Validation:** Developing mechanisms for validating AI responses against source documents.

7.2.3 User Experience Enhancements

Interface Improvements

- **Mobile Optimization:** Ensuring full functionality on smartphones and tablets.
- **Accessibility Compliance:** Meeting WCAG 2.1 AA standards for users with disabilities.
- **Performance Optimization:** Reducing load times and improving responsiveness.

Educational Features

- **Guided Tours:** Interactive tutorials introducing system capabilities and historical context.
- **Sample Queries:** Curated example queries demonstrating interesting historical findings.
- **Contextual Help:** Inline explanations of censorship patterns, historical context, and AI limitations.

7.3 Possibilities for Scalability or Adaptation

7.3.1 Scaling Strategies

Horizontal Scaling

- **Load Balancing:** Distributing queries across multiple backend instances for higher concurrent user capacity.
- **Database Sharding:** Partitioning vector database by time period or collection for improved query performance.
- **CDN Integration:** Using content delivery networks for serving document images and static assets.

Vertical Scaling

- **Hardware Upgrades:** Moving to more powerful GPUs and increased memory as archive grows.
- **Embedding Optimization:** Using dimensionality reduction techniques to decrease storage and computation requirements while maintaining quality.

7.3.2 Adaptation to Other Contexts

Geographic Adaptation

- **Other Spanish Regions:** Applying to archives in Basque Country, Galicia, Madrid, etc. with region-specific historical context.

- **Other Countries:** Adapting for censored archives in other countries (Portugal under Salazar, Latin American dictatorships, Eastern European communist regimes).
- **Language Adaptation:** Modifying embedding models and LLMs for other languages while maintaining architecture.

Temporal Adaptation

- **Earlier Periods:** Extending to Spanish Civil War (1936-1939) or earlier 20th century.
- **Later Periods:** Covering transition to democracy (1975-1982) and democratic era.
- **Contemporary Archives:** Adapting for modern digital-native archives with different characteristics.

Domain Adaptation

- **Different Media Types:**
  - Newspaper archives
  - Film and photography archives
  - Audio recordings (radio, interviews, speeches)
  - Government documents and official records
- **Thematic Focus:**
  - Women's history and feminist movements
  - Labor and worker's movements
  - LGBTQ+ history
  - Regional languages and minority cultures

Institutional Adaptation

- **Smaller Institutions:** Simplified deployment for municipal archives and local historical societies with limited technical capacity.
- **Larger Institutions:** Enterprise deployment for national libraries and major research institutions with extensive collections.
- **Collaborative Networks:** Federation of multiple institutional archives into searchable network while maintaining local control.

7.3.3 Architectural Flexibility

The system's modular architecture enables various adaptation scenarios:

- **Component Independence:** Frontend, backend, preprocessing pipeline, and databases can be independently modified or replaced.

- **API-Driven Design:** Clear API contracts enable alternative frontends (mobile apps, voice interfaces, integration with existing archival systems).
- **Model Agnostic:** Support for multiple LLM providers and embedding models allows optimization for different languages, domains, and cost structures.
- **Containerization:** Docker-based deployment simplifies adaptation to different hosting environments (on-premise servers, cloud platforms, hybrid approaches).

8 Final Reflection

8.1 Critical and Personal Conclusion

This project represents an ambitious attempt to bridge historical scholarship and cutting-edge artificial intelligence within severe time and resource constraints. The five-day intensive development process forced rapid decision-making, pragmatic compromise, and creative problem-solving. The resulting system—while imperfect—demonstrates the viability of using modern AI technologies to democratize access to historically significant but difficult-to-navigate archival materials.

The project's greatest strength lies in its problem-oriented approach. By maintaining close engagement with archivists at the National Library of Catalonia and Universitat Autònoma de Barcelona, the team ensured that technical capabilities addressed real user needs rather than serving as technology demonstrations. The side-by-side comparison feature, for instance, emerged directly from stakeholder conversations and represents a genuine innovation in archival access tools.

The collaborative and interdisciplinary nature of the work proved essential. Team members brought diverse expertise in AI, software engineering, and historical research, creating synergies that enhanced outcomes beyond what any individual could achieve. The division into specialized subgroups (OCR/preprocessing, web application, prompt engineering) enabled efficient parallel development while maintaining integration coherence through regular coordination.

However, the project also revealed significant tensions. The five-day timeline, while fostering focus and intensity, prevented deeper exploration of critical challenges. Censorship detection—arguably the most historically important capability—remains largely unaddressed beyond documentation as future work. The decision to use commercial API (Gemini) rather than local models prioritized immediate functionality over long-term sustainability and institutional autonomy, creating technical debt that stakeholders explicitly identified.

8.2 Evaluation Relative to Initial Objectives

8.2.1 Success Metrics

Primary Objective: Democratizing Access Status: Largely Achieved

The system successfully enables broader access to Radio Barcelona SER archives:

- Semantic search allows users to find relevant documents without knowing exact keywords or archival organization

- Conversational chat interface lowers technical barriers for non-expert users
- Side-by-side comparison addresses the verification gap identified in initial problem statement
- Containerized deployment enables installation at institutions worldwide

However, full democratization requires addressing remaining dependencies on commercial APIs and completing migration to production-ready infrastructure.

Secondary Objective: Revealing Hidden Content Status: Partially Achieved
The system makes censored materials searchable and analyzable:

- Users can query historical content that was previously inaccessible or difficult to navigate
- RAG architecture enables synthesis across multiple documents, revealing patterns not visible in individual documents

However, the critical limitation remains: the system cannot reliably identify what was censored, altered, or removed. Without explicit censorship markers, reconstructing suppressed narratives remains beyond current capabilities. This represents the most significant gap between aspirations and outcomes.

Addressing Stated Problems

- **OCR Difficulty:** Systematic testing identified better-performing solutions, but fundamental OCR challenges persist, particularly for handwritten documents. This problem is mitigated but not solved.
- **Navigation Challenges:** Semantic search and conversational interface dramatically improve browsability. This problem is substantially solved within the scope of processed documents.
- **Verification Gap:** Side-by-side comparison directly addresses the lack of tools for comparing OCR'd text with originals. This problem is solved for included documents.
- **Access Barriers:** Digital system eliminates physical access requirements. This problem is solved, though digital access creates new barriers (internet connectivity, digital literacy).

8.2.2 Unexpected Outcomes

Several outcomes exceeded or differed from initial expectations:

- **Stakeholder Engagement Value:** The transformative impact of direct archivist collaboration exceeded expectations. Initial technical plans were substantially reshaped by stakeholder insights.
- **RAG Architecture Learning:** Team members new to RAG systems developed significant competency within five days, demonstrating feasibility of intensive skill development when properly structured.

- **Time Constraint Benefits:** While limiting depth, the five-day deadline prevented scope creep and perfectionism, forcing focus on essential functionality.
- **Prompt Engineering Complexity:** The difficulty of minimizing hallucinations through prompt engineering alone proved greater than anticipated, requiring dedicated subgroup effort.

8.3 Limitations of the Project

8.3.1 Scope Limitations

- **Temporal Scope:** Focus on 1945 alone provides proof of concept but limits generalization. Patterns observed may not represent other years or periods.
- **Archive Coverage:** Radio Barcelona SER represents one specific type of archival material. Newspapers, government documents, and personal correspondence may present different challenges requiring adaptation.
- **Geographic Specificity:** System is optimized for Barcelona and Catalonia. Adaptation to other regions requires consideration of linguistic, cultural, and historical differences.
- **Representative Sample Uncertainty:** Unknown whether 1945 subset adequately represents challenges present in full archive.

8.3.2 Technical Limitations

- **OCR Quality Ceiling:** Fundamental OCR limitations remain unchanged. Handwritten documents, degraded scans, and stamp-obscured text continue to produce poor results.
- **Censorship Blindness:** Critical inability to detect censored, altered, or removed content undermines historical interpretation capabilities.
- **Hallucination Risk:** Despite prompt engineering, LLMs can generate incorrect information. Users require historical expertise to critically evaluate responses.
- **Computational Requirements:** Estimated GPU memory requirements (18GB) for full archive may be underestimated. Actual scaling needs remain uncertain.
- **API Dependency:** Current reliance on commercial Gemini API creates costs, dependencies, and potential service discontinuation risks.

8.3.3 Methodological Limitations

- **Evaluation Challenges:** Assessing historical accuracy and interpretation quality is inherently subjective. No quantitative metrics fully capture system success.
- **User Testing Gap:** Limited time prevented extensive user testing with actual researchers, students, and public users. Real-world usability may differ from anticipated.

- **Bias Unexamined:** Insufficient analysis of how AI model biases might affect interpretation of historical materials, particularly propaganda and censored content.
- **Sustainability Planning:** Development focused on implementation rather than long-term maintenance, governance, and funding models.

8.3.4 Resource Limitations

- **Time Constraints:** Five days permitted proof of concept but prevented depth in critical areas (censorship detection, local model deployment, comprehensive testing).
- **Funding Limitations:** Reliance on free-tier services and existing hardware constrained scale and capability.
- **Expertise Gaps:** While interdisciplinary collaboration was valuable, team lacked deep expertise in some critical areas (paleography, censorship studies, production-scale ML deployment).

8.4 Open Questions

8.4.1 Technical Open Questions

- **Censorship Detection:** Can computer vision and NLP techniques reliably identify censorship markers, altered content, and patterns of suppression in historical documents? What accuracy levels are achievable?
- **Embedding Strategy:** Does structure-preserving Markdown embedding outperform pure text embedding for historical documents? Under what conditions is each approach optimal?
- **Local Model Viability:** Can open-source local LLMs (LLaMA, Mistral) match Gemini performance on Spanish/Catalan historical queries when fine-tuned? What computational requirements does this impose?
- **Scalability Validation:** Are GPU memory estimates accurate for full archive? What other bottlenecks emerge at scale?
- **Multi-Modal Integration:** How can image-based analysis be integrated with text-based search for documents containing significant visual elements (photographs, illustrated propaganda)?

8.4.2 Methodological Open Questions

- **Evaluation Frameworks:** What metrics and methodologies can meaningfully assess accuracy and value of AI-assisted historical research tools?
- **Ground Truth Establishment:** How can "correct" answers be established for historical queries when interpretation is inherently subjective and contested?
- **Bias Detection:** What systematic approaches can identify and mitigate AI model biases when applied to propaganda and censored materials?

- **Human-AI Collaboration:** What division of labor between AI systems and human researchers produces optimal historical research outcomes?

8.4.3 Sustainability Open Questions

- **Funding Models:** What sustainable funding models support ongoing maintenance, development, and scaling? Should access be free, subscription-based, or institutionally funded?
- **Governance:** Who should control system development, content inclusion decisions, and access policies? How should stakeholders be represented in governance?
- **Maintenance Requirements:** What realistic staffing and budget are required for long-term operation? Can volunteer or academic labor sustainably maintain the system?
- **Update Strategies:** As AI technology evolves, how frequently should embedding models and LLMs be updated? How to balance stability with improvement?

8.4.4 Social and Ethical Open Questions

- **Access Equity:** Does digital access genuinely democratize historical knowledge, or does it create new barriers (digital literacy, internet access) that reproduce existing inequalities?
- **Authority and Trust:** How do users distinguish between reliable AI responses and hallucinations? What responsibility do system creators bear for potential misinterpretations?
- **Historical Interpretation:** Does AI synthesis of archival materials inappropriately simplify complex historical questions? Does it encourage over-reliance on computational tools at the expense of critical human analysis?
- **Cultural Heritage Ownership:** Who owns and controls digitized cultural heritage? How should international access be balanced with national or regional ownership?
- **Propaganda Amplification:** Does making censored propaganda easily accessible risk amplifying harmful historical narratives without adequate contextualization?

8.4.5 Domain-Specific Open Questions

- **Censorship Patterns:** What systematic patterns of censorship existed in 1945 Radio Barcelona materials? How did these evolve over time and compare to other media?
- **Missing Content:** What types of content were most heavily censored or suppressed? Can we reconstruct likely topics or narratives that were eliminated?
- **Propaganda Mechanisms:** How did propaganda operate in Radio Barcelona broadcasts? What rhetorical and content strategies were employed?

- **Historical Significance:** What new historical insights emerge from broader access to these materials? How does radio content compare to newspaper censorship patterns?

8.5 Future Directions

The project establishes a foundation with multiple potential development paths:

8.5.1 Immediate Next Steps

- Resolve technical debt (local LLM migration, production database deployment, PDF embedding)
- Conduct user testing with historians, students, and public users
- Develop comprehensive documentation and training materials
- Establish governance and maintenance structures with National Library of Catalonia

8.5.2 Medium-Term Developments

- Expand temporal coverage to broader periods (1940s, eventually 1939-1975)
- Develop censorship detection capabilities through interdisciplinary research
- Implement advanced features (filtering, export, multilingual interface)
- Evaluate and refine embedding strategies based on usage patterns

8.5.3 Long-Term Vision

- Create federated network of censored archives from multiple institutions and countries
- Develop specialized AI models fine-tuned for historical Spanish/Catalan documents and censorship analysis
- Establish the system as a case study and template for archival AI applications globally
- Contribute to broader understanding of how AI can responsibly augment humanities research

8.6 Concluding Thoughts

This project demonstrates both the promise and limitations of applying modern AI technologies to historical archives. Within five intensive days, a small team created a functional system that meaningfully addresses real archival access challenges. The successful implementation proves that sophisticated RAG-based systems can be built rapidly when properly scoped and problem-oriented.

However, the project also reveals that technical capability alone is insufficient. The most critical challenge—detecting and interpreting censorship—remains unsolved, highlighting the gap between what AI can currently achieve and what historians most need. This gap suggests that the future of archival AI lies not in fully automated analysis but in thoughtful human-AI collaboration, where computational tools augment rather than replace expert human judgment.

The system's value ultimately depends on continued development, sustained institutional commitment, and careful attention to ethical considerations around access, interpretation, and historical responsibility. Success will be measured not by technical sophistication but by whether the system genuinely enables new historical understanding and broader engagement with Catalonia's complex 20th-century history.

As Franco-era Spain recedes further into the past, the urgency of preserving and understanding this period grows. Projects like this represent not merely technical exercises but contributions to collective memory and historical reckoning. The question is not whether AI can perfectly analyze historical archives—it cannot—but whether it can help humans ask better questions, find relevant evidence more efficiently, and engage more critically with difficult periods of history. On this measure, the Barcelona Archives Project offers genuine promise.

Acknowledgments

This project would not have been possible without the collaboration and support of:

- Archivists at the National Library of Catalonia for providing access to materials, historical context, and invaluable domain expertise
- Collaborators at Universitat Autònoma de Barcelona for their insights into information retrieval challenges and archival practices
- The workshop organizing team for creating an intensive and productive learning environment
- All team members who contributed their diverse expertise and worked tirelessly across five intensive days