DECODING THE SILENCE

Lucia Rubio, Rubén Oliver, Marta Haik, Arthur Auclair, Diana Efimova, Tom Hillbrunner, Giuseppe Stefano Blanco, Nojus Kondrotas, Precious Idoro, Erikas Zaura, and Corentin Brandam

PROJECT OVERVIEW

Topic and Main Objective

This project focuses on the analysis of radio scripts from Radio Barcelona in order to identify the presence and patterns of censorship during the Franco dictatorship. Through the systematic study of these scripts, our main objective is to detect censorship marks, interventions, and silences imposed on radio content, and to classify them according to different thematic categories.

By examining the scripts, we aim to determine which topics were more heavily censored, how censorship was applied, and to what degree each document was affected. These themes include, among others, politics, social issues, morality, gender roles, religion, and cultural identity.

To support this analysis, we have developed an AI-based system that processes the digitised documents and organises them by thematic category. The images of the radio scripts are visually separated according to these themes, and a colour-coded system is used to indicate the level of censorship: documents that show clear evidence of censorship, documents with no visible censorship, and documents where censorship cannot be conclusively determined.

Through this approach, the project combines historical research and artificial intelligence to make censorship visible, measurable, and accessible, offering a new way to explore how control over radio broadcasting shaped public discourse during the dictatorship.

Problem or Guiding Question

This project is guided by three central questions that address the nature, mechanisms, and motivations of radio censorship during the Franco dictatorship.

1. **What was hidden or erased?**
   By analysing the radio scripts, the project seeks to identify which topics, expressions, or narratives were systematically removed, altered, or silenced. This includes not only explicit political content, but also references to social conflict, gender roles, cultural identity, and everyday realities that contradicted the official discourse of the regime.

2. **What patterns exist?**
   Rather than treating censorship as a series of isolated interventions, the project examines whether recurring patterns can be identified across time, programmes, and themes. By comparing scripts and categorising them thematically, we aim to reveal regularities in how censorship was applied, which topics were more frequently targeted, and whether certain forms of language or formats were more vulnerable to intervention.

3. **Why were they hidden or erased?**
   This question addresses the underlying logic of censorship. The project explores how censorship functioned as a tool to maintain political control, enforce moral norms, and shape social behaviour. By situating the censored material within its historical and ideological context, we seek to understand the broader objectives of the regime and the reasons certain voices, ideas, or realities were considered unacceptable for public broadcast.

Together, these guiding questions allow the project to move beyond identifying censorship as a phenomenon, towards understanding it as a structured system of power that shaped radio content and public discourse over time.

Context in Which the Project is Situated.

To understand this project, it is essential to place it within its historical context. After the end of the Spanish Civil War in 1939, the regime led by Francisco Franco faced the challenge of consolidating its power over a defeated, impoverished, and deeply divided society. In this context, censorship became an essential tool to eliminate any remaining traces of the political and cultural pluralism of the Second Spanish Republic. Radio, due to its ability to reach the population immediately and on a massive scale, was one of the media most closely monitored and controlled by the state.

From a legal standpoint, the control of information was institutionalised through the Press Law of 22 April 1938, approved while the war was still ongoing. This law established that the media had to serve the "national interests," which in practice meant mandatory adherence to the ideology of the new regime. Although the law primarily referred to the printed press, its principles were directly applied to radio broadcasting, which came to be regarded as a first-order political instrument.

One of the key historical developments that decisively shaped radio was the creation and consolidation of Radio Nacional de España (RNE). Founded in 1937, RNE became, after the war, the only broadcaster authorised to disseminate political information. From the 1940s onwards, all private radio stations were required to connect several times a day to RNE to broadcast the official news bulletins, popularly known as the Bulletin. This system established a true information monopoly: private stations were allowed to entertain, but not to inform.

The 1940s, marked by economic autarky and the regime's international isolation, represented the harshest period of censorship. Radio content was subject to prior approval, and any ideological deviation could result in fines, temporary closures, or the permanent withdrawal of broadcasting licences. Censorship did not focus solely on politics, but also extended to moral and cultural matters, reinforcing the alliance between the regime and the Catholic Church.

The most strictly censored topics were, first and foremost, political ones: references to the Republic, exile, outlawed political parties, or any form of opposition were absolutely prohibited. Social issues were also closely monitored; strikes, poverty, inequality, or labour conflicts could not be mentioned, as they contradicted the image of social harmony the regime sought to project. Moral censorship affected everything related to sexuality, family roles, divorce, and female behaviour, always from a deeply conservative perspective. Finally, cultural and linguistic diversity was harshly repressed: the use of

Catalan, Basque, or Galician was banned from radio broadcasting for decades, reinforcing the regime's policy of cultural uniformity.

Within this framework, radio also played a fundamental role in shaping gender roles. Women were allowed to participate in radio broadcasting, but under strict limitations. Female announcers existed from the early years of the dictatorship, yet their roles were confined to programmes considered "feminine": domestic advice, religious content, light music, or children's programmes. Radio women were expected to embody the regime's ideal of femininity submissive, maternal, discreet, and removed from public debate. They were barred from political news and opinion programmes, which were reserved exclusively for male voices. Even their tone of voice and manner of expression were subject to supervision, ensuring that they conveyed gentleness, obedience, and moral propriety.

From the 1950s onwards, as Spain gradually emerged from international isolation and signed agreements with the United States, the regime initiated a slow and limited opening. Radio programming began to diversify through game shows, serial dramas, and music programmes, slightly softening the perception of control, although censorship remained fully in force. The true turning point came in 1966 with the approval of the Press and Printing Law promoted by Manuel Fraga. This law formally abolished prior censorship, replacing it with a system of post-publication responsibility. In practice, however, radio continued to operate under strict controls: economic and administrative sanctions encouraged self-censorship, and RNE's monopoly over political information remained unchallenged.

During the final years of the dictatorship, particularly in the early 1970s, tentative spaces for greater cultural and social openness began to appear. Some programmes addressed youth culture, music, or science with greater freedom, and women slowly gained visibility in more diverse formats. Nevertheless, censorship did not truly disappear until Franco's death in 1975 and the beginning of the democratic Transition, when freedom of expression was progressively recognised and state control over radio broadcasting finally came to an end.

WORKING WITH ARCHIVES

Criteria for Selection and Organization of The Archive and Data

Because our corpus consisted exclusively of radio scripts from the year 1942, our selection process was shaped both by what the archive offered and by what the project required. The first criterion was simply readability. The scans varied in quality, and some pages were very faded or had slight distortions from the original paper. We focused on including documents where the typography, handwritten notes, and layout were sufficiently visible for both human analysis and AI processing. This also meant trying to preserve as much variety as possible in the scripts so we could see which themes appeared across the documents and how they might relate to censorship practices (which later contributed to the selection of the main cluster's edges).

Organising the data became easier once all pages had been converted into image files. We numbered them based on their order in the digitised folders, but our real organisational structure only emerged once the AI began generating embeddings. These embeddings created a semantic and visual space in which pages positioned themselves based on similarity. Instead of imposing a predefined hierarchy, we used this structure to observe natural clusters, which we later aligned with the thematic categories we were studying, such as religion, entertainment, social topics, and possible references to Catalan identity.

Critical Decisions, Limitations, or Tensions Encountered When Working With Archival Materials

Working with the archival documents forced us to confront several practical limitations and methodological tensions. One of the most important decisions involved distinguishing censorship from normal editorial activity. Since these documents were working materials from a radio station, not every correction or annotation carried historical weight. Some pencil marks looked like quick author edits, while others seemed more forceful or externally imposed. Without clear metadata, we had to make interpretative judgments, which meant accepting a certain level of uncertainty in how we classified these interventions.

Another challenge came from the physical condition of the pages. Some scans were blurry or uneven, making it harder for the AI to detect and interpret visual cues. In a few cases, the layout itself confused the model, especially when parts of pages were cut off or handwritten notes overlapped with typed text. These imperfections are typical in archives, but they highlighted how fragile historical sources can be, and how dependent digital analysis is on the quality of the material it receives.

A deeper tension emerged from the time constraints of the archive itself. Because our set only included documents from 1942, we were analysing censorship within an isolated moment rather than across multiple years. This prevented us from identifying long-term evolutions or comparing phases of stricter and looser control. At the same time, focusing on a single year forced us to engage more closely with the specific atmosphere of early-Francoist radio production, where political content was almost entirely suppressed and moral messaging dominated the scripts.

Finally, we confronted the limits of our AI-assisted methodology. The multimodal embeddings sometimes prioritised visual similarity over thematic meaning, leading to clusters that require manual reinterpretation (which couldn't really be done during the short sprint). For example, when searching for "divorce", the search engine doesn't look for "ending the marriage" or

“parting ways”, but rather this exact wording. This created a constant back-and-forth between what the model detected and what we, as human readers, believed was historically plausible.

Together, these decisions and tensions shaped not only the structure of our dataset but also our understanding of how censorship becomes visible or remains hidden, within archival material.

PROCESS DEVELOPMENT

Main Stages of The Process

The project followed an iterative development process, with each stage informed by the limitations discovered in the previous one.

Our first engagement with the Radio Barcelona archive began with a series of meetings initiated in November. However, we only fully understood the nature and complexity of the material we would be working with after our first in-person meeting with Tomàs Fàbregat Anglès, Director of the Library at the Autonomous University of Barcelona, and Rosa Cabezas García, Administrative Officer responsible for the archive.

Initial exploration began with conventional approaches. Traditional OCR pipelines were tested first but proved inadequate—the material quality (photocopies, handwritten annotations, mixed layouts) resulted in unreliable text extraction. This failure motivated the shift toward vision-language models that process pages as images directly, bypassing text extraction entirely.

The technical aspects of our approach will be examined in greater detail in the following section. What we aim to focus on here, instead, is the interpretative framework through which we chose to approach this assignment.

With the embedding pipeline established, the first functional output was a semantic search engine. Early testing confirmed that the approach worked—queries returned semantically relevant results, and the system handled cross-lingual queries effectively. However, search alone provided no way to explore the archive's overall structure or identify patterns across thousands of documents.

This led to the development of the cluster map as a complementary visualization tool. Defining meaningful thematic clusters required several iterations—initial cluster descriptions were either too broad or too narrow, and the spatial layout needed tuning to produce interpretable groupings.

For censorship classification, the first approach used text-based classification queries—attempting to describe what censored documents look like in natural language. This performed

poorly because censorship manifests in many visual and contextual dimensions that are difficult to capture in a single prompt. The shift to prototype learning (using example images rather than text descriptions) proved far more effective, as the model could learn directly from visual examples.

Challenges Encountered and How They Were Addressed

OCR failure on archival material: The heterogeneous nature of the source documents made text extraction unreliable. Addressed by adopting a purely visual embedding approach.

Flat search results lack context: A ranked list of results provided no sense of thematic structure. Addressed by developing the cluster map visualization.

Text prompts cannot capture visual censorship patterns: Describing censorship in words failed to capture its visual manifestations (stamps, crossed-out text, handwritten annotations). Addressed by switching to prototype learning with curated example images.

Binary classification too rigid: Initial attempts at censored/uncensored classification produced too many uncertain cases. Addressed by introducing an explicit "Uncertain" category with a confidence threshold.

Cluster definitions required iteration: Early thematic clusters were poorly calibrated. Refined through testing with sample queries and adjusting text descriptions until clusters produced meaningful groupings.

Image resolution trade-offs: Higher resolution improved detail recognition but increased processing time and memory usage. Addressed by testing different resolutions and settling on a balance (~1 megapixel per page).

Ambiguity in defining censorship: One of the main challenges was determining what could genuinely be considered censorship and what could not. The documents were filled with corrections and handwritten marks, yet attributing these with certainty either to government censorship or to simple self-corrections was often nearly impossible.

We ultimately concluded that addressing this ambiguity required training the model through human supervision—explicitly confirming what should be classified as censorship and what should not, and guiding the system toward a deeper, context-aware interpretation. This proved essential, as censorship was frequently embedded in what was absent from the texts, making it particularly difficult to detect without careful interpretative analysis.

This reinforced the need to treat censorship not as a purely visual or textual feature, but as an interpretative phenomenon emerging from historical, political, and contextual cues.

Key Learnings Emerging From the Process

Failed approaches are informative. The OCR failure and the ineffective text-based classification both pointed toward better solutions—each dead end clarified what the problem actually required.

Visual examples outperform verbal descriptions for pattern recognition. Prototype learning succeeded where prompt engineering failed, suggesting that some concepts are easier to demonstrate than to describe.

Complementary interfaces serve different research modes. Search supports targeted inquiry; the cluster map supports exploratory discovery. Building both on the same embedding foundation maximized the value of the initial computation.

Explicit uncertainty is more useful than forced classification. Acknowledging what the system cannot confidently determine proved more valuable than producing unreliable binary labels.

TECHNOLOGICAL DIMENSION

Technologies Used

The project is built on Qwen3-VL-Embedding-8B, a state-of-the-art vision-language model that currently ranks as the top-performing architecture for multimodal embeddings. This model processes each archival page as a high-resolution image and generates a 4096-dimensional embedding vector that captures both visual features (layout, stamps, handwritten annotations, red censor marks) and semantic content (the meaning of visible text) in a unified representation. Crucially, this approach bypasses traditional OCR pipelines entirely—the model understands page content directly from the visual input.

The generated embeddings are stored in LanceDB, a vector database optimized for cosine similarity search. On top of this foundation, two complementary systems were built: a semantic search engine and an interactive cluster map visualization. For improved search accuracy, an optional second-stage reranker (Qwen3-VL-Reranker-2B) can rescore results with live progress feedback.

The entire pipeline is written in Python and runs locally on Linux-based systems with AMD hardware, requiring approximately 18 GB of VRAM. Pre-Processing speed averages 0.8 seconds per page, making it feasible to index large archives on a single workstation.

Technology Process

The technological workflow consists of three stages: embedding generation, search functionality, and cluster visualization.

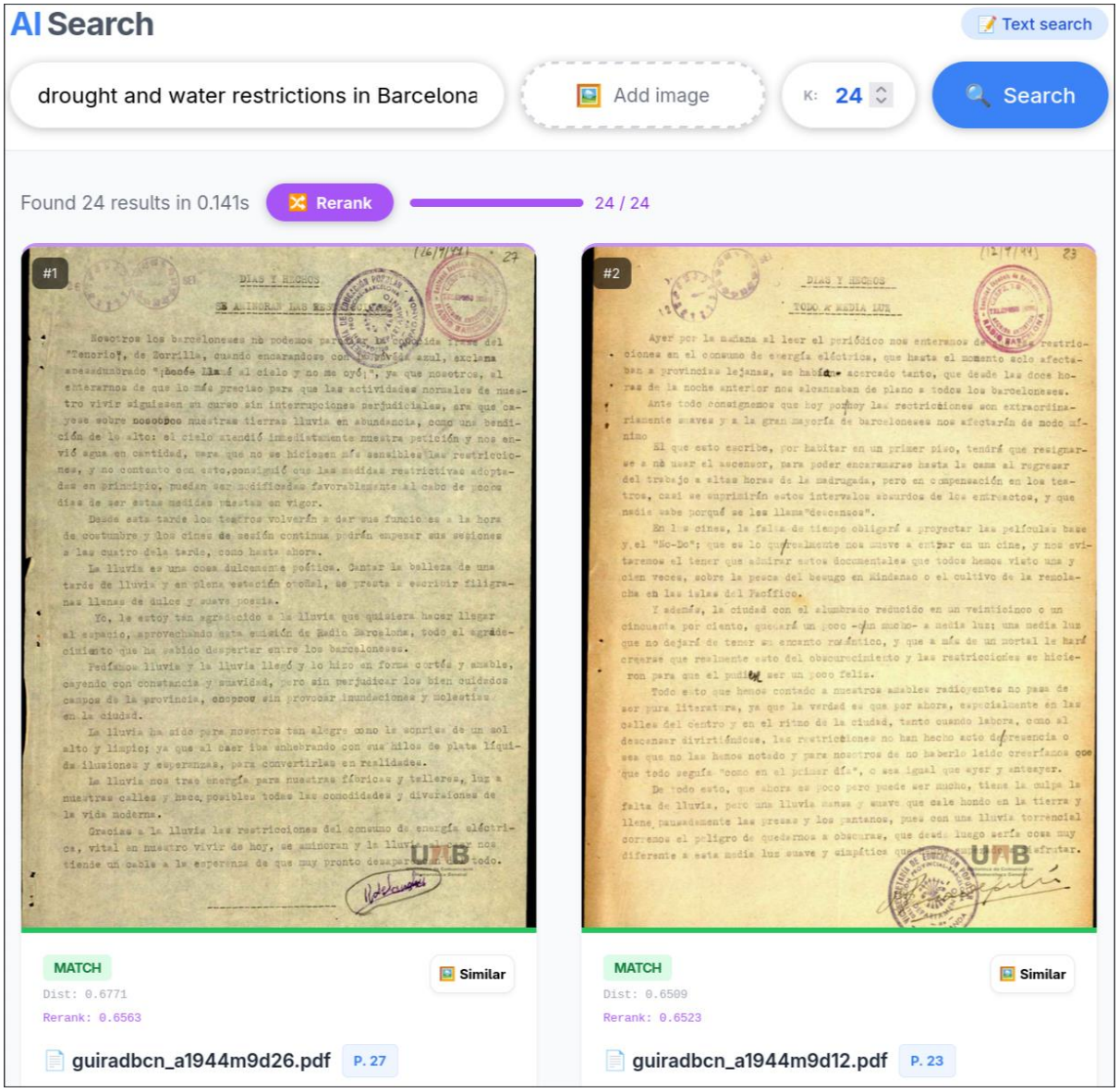
Stage 1: Embedding Generation (Foundation)

The preprocessing pipeline converts PDF documents into individual page images, resizes them to a target resolution (~1 megapixel), and passes each image through the Qwen3-VL-Embedding-8B model. The resulting 4096-dimensional vectors are stored in the LanceDB vector database alongside metadata (source document, page number). This embedding computation forms the foundation upon which both the search engine and the cluster map are built.

Stage 2: Semantic Search Engine

The search engine enables users to query the archive through multiple input modes: text queries, image uploads, or a combination of both. When a user submits a query, it is converted into an embedding using the same model, and the system retrieves the most similar pages via cosine similarity search—returning results in less than one second.

For enhanced precision, an optional reranking step uses the Qwen3-VL-Reranker-2B model to rescore the initial results. This second-stage model evaluates each candidate page against the query and provides a refined relevance score, with progress displayed in real-time through the web interface.



[Screenshot: Search interface with example query]

Stage 3: Cluster Map Visualization

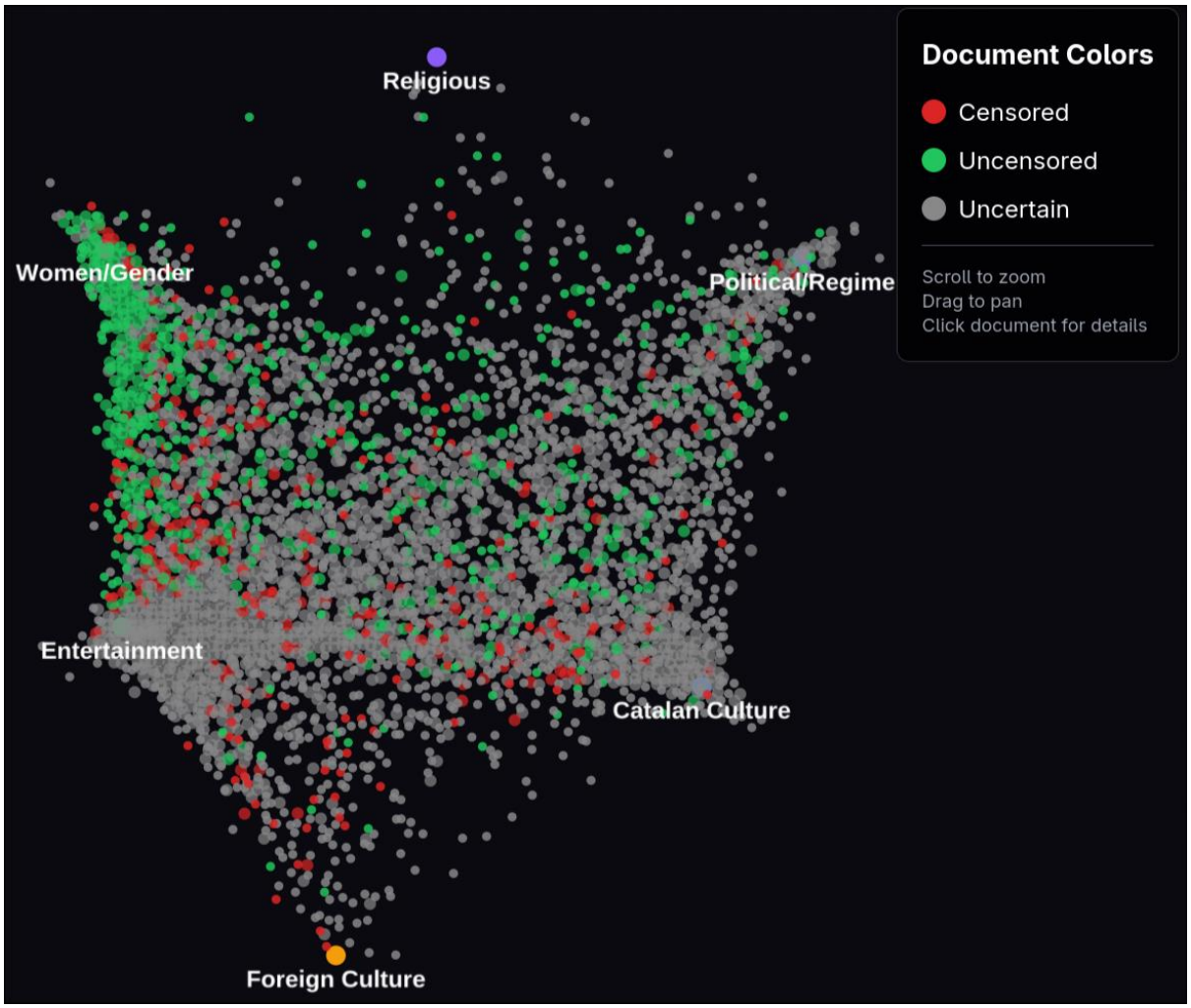
Building on the same embedding foundation, the cluster map projects all documents onto a 2D interactive visualization. This system uses two types of clusters:

- Position clusters define thematic user-defined categories (such as Religion, Catalan Culture, Entertainment, Political/Regime content). Each cluster is represented by a text description that is embedded into the same vector space. A neural network projection head is trained to map each document's similarity profile to these clusters into 2D coordinates, preserving both local neighborhood structure and global thematic relationships.

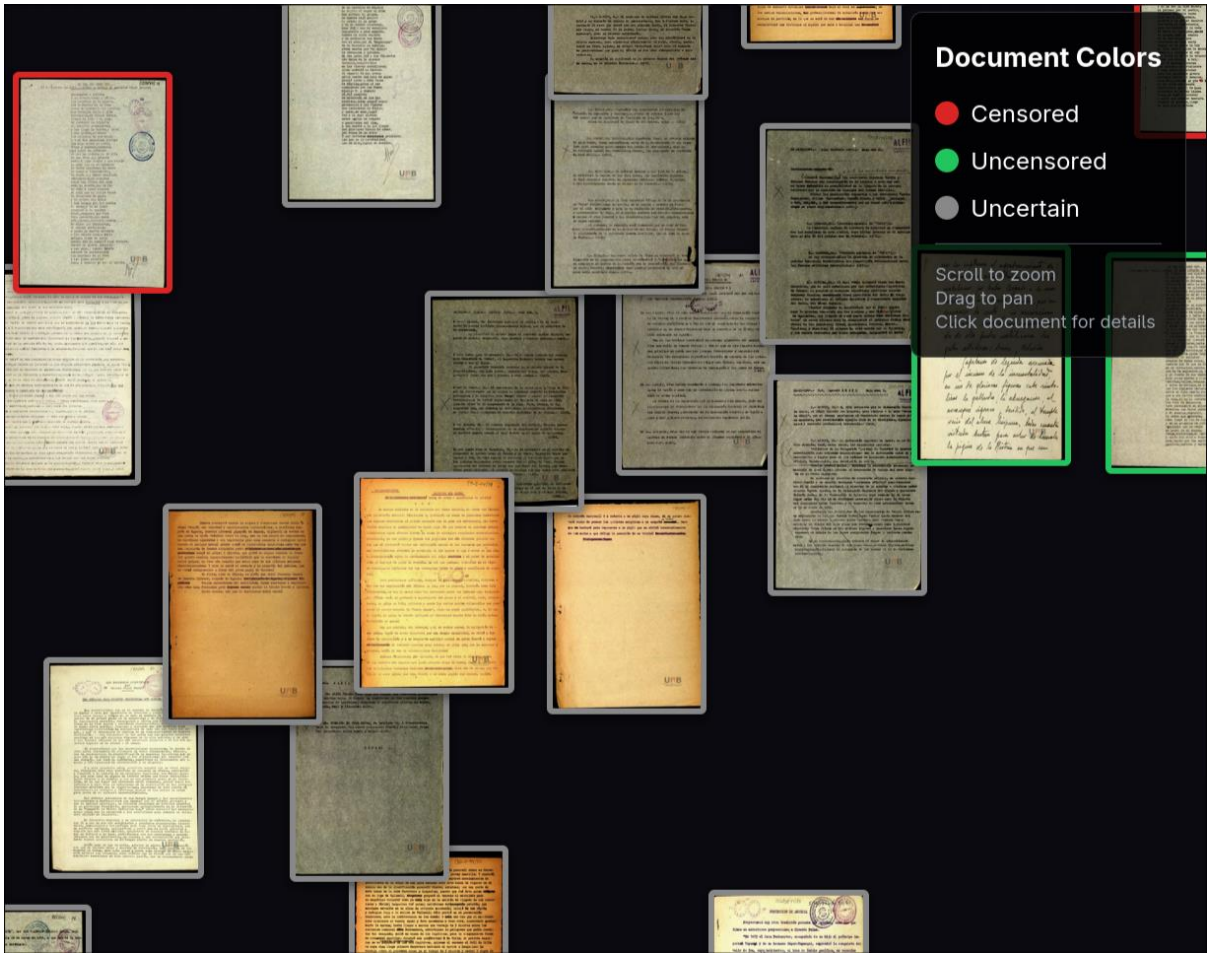
- Color clusters determine document coloring through prototype learning. Approximately 30 hand-selected example pages for each category (censored and uncensored) were embedded to create prototype vectors. Each document's color is determined by its maximum similarity to these prototypes: red indicates high likelihood of censorship, green indicates uncensored

content, and gray marks documents where the model cannot confidently determine a classification (below a confidence threshold).

The resulting interactive map allows users to explore the archive spatially. At low zoom levels, a density heatmap provides an overview of document distribution. Zooming in reveals individual documents as colored points, and at the closest zoom level, page thumbnails become visible for direct inspection.



[Screenshot: Cluster map overview showing thematic regions with colored document points]



[Screenshot: Cluster map zoomed in, showing individual documents with censorship coloring and visible thumbnails]

Role of Technology in Meaning-making

Technology plays a central role in how meaning is produced in this project by enabling a unified multimodal representation of the archive. Rather than treating text and image as separate channels requiring different tools (OCR for text, image classifiers for visuals), the embedding model captures both dimensions simultaneously. This allows censorship to be approached as a phenomenon that operates at multiple levels: semantic alterations in discourse and material traces such as stamps, crossed-out passages, and handwritten annotations.

The embedding space serves as a shared foundation for two complementary modes of meaning-making. The search engine supports targeted inquiry—users can formulate specific questions and retrieve relevant pages directly. The cluster map supports exploratory discovery—users can navigate thematic regions and observe how documents relate to one another across the archive.

Importantly, the system does not directly classify documents as "censored" or "uncensored" in an authoritative sense. The prototype learning approach highlights patterns of similarity to known examples, surfacing pages that warrant closer human inspection. Meaning therefore

emerges from the interaction between computational pattern detection and human interpretative judgment.

Level of Dependence or Autonomy in Relation to the Tools

The project demonstrates significant dependence on the vision-language model for large-scale pattern detection. Without the embedding computation, neither the sub-second search capability nor the spatial visualization would be possible at this scale. The model's ability to jointly process visual and textual information is essential to capturing the materiality of the archive.

However, the system remains far from autonomous. Human expertise was required at multiple stages:

- Defining the thematic position clusters and writing their text descriptions
- Selecting the reference pages for censored and uncensored prototypes
- Setting the confidence threshold that determines when a document is marked as "Uncertain"
- Interpreting the resulting clusters and search results in their historical context

The tools function as exploratory aids that surface patterns for human analysis rather than as authoritative classifiers. Crucially, the prototype learning approach requires only a small number of manually annotated examples to enable automatic classification across thousands of documents—a significant efficiency gain over manual inspection. However, this design choice also reflects the inherent ambiguity of the source material: distinguishing official censorship from editorial corrections, self-censorship, or authorial revisions often requires contextual judgment that cannot be fully automated.

Technical Constraints, Biases, or lImitations Identified

Several limitations constrain the project's analytical reach. They include:

Absence of ground truth: No labeled dataset exists that definitively identifies which pages were censored. A rigorous evaluation would require a substantial corpus of pages labeled by domain experts (e.g., censored, uncensored, unclassifiable), split into training, validation, and test sets. The training set would be used to compute prototype embeddings, the validation set to tune hyperparameters such as the confidence threshold, and the held-out test set to objectively measure classification performance. Without this methodology, the current system lacks an objective metric for optimization—prototype selection and threshold settings remain based on informal judgment rather than measurable accuracy.

Moreover, the boundaries between censorship categories are inherently fluid—distinguishing official censor interventions from self-censorship, editorial corrections, or authorial revisions is often ambiguous even for domain experts. Some cases are visually obvious (large red stamps, crossed-out

paragraphs), while subtle alterations remain difficult to classify with confidence. This ambiguity explains the substantial number of gray ("Uncertain") documents in the visualization.

Model training context: The Qwen3-VL model was trained on general web data, not on historical radio archives or the political and cultural specifics of Francoist Spain. It may fail to recognize subtle, context-dependent forms of censorship or misinterpret visual markers that held specific meaning in this historical period.

Multimodal ambiguity: While the unified embedding captures both visual and semantic features, it is not always possible to determine which modality drives a particular similarity or classification decision. A document might appear similar to censored examples due to visual layout rather than content, or vice versa.

Representation limits: The 2D cluster map and color-coded classifications necessarily abstract complex historical processes into simplified spatial and chromatic representations. While useful for exploration and pattern detection, these visualizations risk flattening nuance and should be understood as starting points for deeper inquiry rather than definitive analytical conclusions.

Despite these constraints, the system successfully enables a mode of archival exploration that would be impractical through manual inspection alone, surfacing candidate pages and thematic patterns that merit closer scholarly attention. Importantly, the architecture is purely retrieval-based: users only ever see original archival documents, never generated content. This eliminates the risk of hallucination that affects generative AI systems and ensures that the historical integrity of the source material is preserved throughout the analysis.

METHODOLOGIES AND TOOLS

Methodologies Applied

At its core, the methodology is exploratory. Rather than starting from fixed hypotheses about where censorship would appear, the project allows patterns to emerge from the data itself. The use of multimodal embeddings and spatial clustering enables a form of “distant reading” of the archive, where relationships between documents can be observed at scale before being interpreted historically. This exploratory stance is particularly appropriate given the ambiguous nature of censorship traces, which are not always explicitly marked.

The project is also experimental, as it applies vision-language models to historical radio archives—something for which these tools were not originally designed. The decision to bypass OCR and rely instead on direct visual-semantic embeddings constitutes a methodological experiment in itself, testing whether censorship can be detected not only

through textual content but also through material features such as stamps, crossed-out paragraphs and handwritten notes.

An analytical methodology complements this exploratory phase. Once clusters and search results are produced, they can be examined critically through historical and contextual analysis. Documents surfaced by the system are not treated as final evidence, but as candidates for closer reading. Human interpretation is essential to distinguish official censorship from editorial correction, self-censorship, or routine production notes. In this sense, the AI functions as an analytical filter rather than as a final decision-maker.

Finally, the project is inherently collaborative, operating at the intersection of humanities and computer science. Historical knowledge informs the definition of thematic clusters, the selection of prototype examples, and the interpretation of results, while technical expertise shapes the embedding pipeline, database structure, and visualization tools. This collaboration ensures that methodological decisions are informed both by computational feasibility and by historical relevance.

Relationship Between Methodology, Objectives, and Outcomes

The chosen methodologies are closely aligned with the project's central objectives: to identify censorship, reveal patterns across the archive, and understand the logic behind what was silenced.

The exploratory and experimental methods directly support the objective of making censorship visible. By projecting thousands of pages into a shared embedding space and visualizing them spatially, the project reveals concentrations of censorship that would remain invisible through traditional close reading alone. The analytical methodology ensures that these visual patterns are grounded in historical interpretation, preventing purely technical readings detached from context. Similarly, the combination of semantic search and clustering supports the objective of identifying patterns. Rather than focusing on isolated documents, the methodology enables comparison across themes, formats, and degrees of censorship, highlighting regularities in how the Francoist regime intervened in radio content.

Finally, the collaborative and interpretative dimensions of the methodology address the question of why certain content was hidden or erased. By situating computational results within the political, moral, and cultural framework of the dictatorship, the project connects technical outputs to broader structures of power and control.

Overall, the methodological approach ensures that outcomes are not limited to technical demonstrations, but contribute meaningfully to historical inquiry. The tools do not replace

archival interpretation; instead, they expand its scale, surface hidden relationships, and open new paths for understanding how censorship operated as a systematic and institutional practice.

PROPOSED USE AND REAL-WORLD APPLICATION

Projection of the Project Beyond the Classroom and Potential Contexts of Application

- **Educational**: It can serve as an interactive teaching tool in universities and secondary schools to visualize the "invisible" mechanics of a dictatorship.
- Cultural & Archival: Libraries and national archives (like the Arxiu Nacional de Catalunya) could integrate this AI to automatically catalog and index thousands of unprocessed boxes of radio scripts.
- Professional (Journalism & Media): Modern media outlets could use the tool to create data-driven documentaries or investigative pieces on the evolution of freedom of speech in Spain.Social: It contributes to "Democratic Memory" (Memoria Democrática), helping the public understand how daily language and culture were shaped by institutional fear.

Intended Users

- Archivists (Professional & Cultural Context): The AI functions as an automated indexing system for large-scale historical collections. By identifying government stamps and categorizing censorship types (social, political, religious), it eliminates the need for manual review of thousands of fragile documents. This facilitates the rapid digitization and cataloging of archives, such as the Arxiu Nacional de Catalunya, making previously "hidden" documents searchable for the first time.
- Students (Educational & Social Context): The platform serves as an interactive research tool for students of History, Journalism, and Data Science. It enables "distant reading" of primary sources, allowing users to observe shifts in social taboos and state-mandated language across a chronological and geographical map of Barcelona. This converts abstract historical concepts into a tangible, digital experience that promotes engagement with "Democratic Memory."
- Teachers (Pedagogical Context): Educators can utilize the tool as a dynamic classroom resource to provide empirical evidence of how a dictatorship functions. By filtering for specific categories—such as the censorship of women's content or religious critiques—teachers can demonstrate the regime's priorities with hard data. It also provides a

practical case study for discussing media ethics and the evolution of freedom of expression.

Real-World Value and Impact

The primary impact of the project lies in the quantification of suppression. By extracting the "hidden layer" of corrected or deleted text, the AI preserves historical data that was originally intended to be erased. The interactive visualization maps the geographical and thematic obsession of the Francoist regime, moving censorship from an anecdotal historical fact to a measurable dataset. Ultimately, the project provides a scalable framework for digital humanities, ensuring that the mechanics of institutional censorship remain visible and analyzed in a modern, democratic context.

STEPS TOWARD IMPLEMENTATION

In order for the tool to encompass the complete archive, it must index documents across all available years rather than a single year covered by the current prototype representative implementation. The full implementation requires substantial computational infrastructure for locally deploying a visual-semantic model (such as Qwen3-VL-Embedding) for document page analysis, efficient embedding generation and processing pipelines, as well as an adequate storage system accommodating both raw digitized materials and their vector representations. Additional computational capacity must be dedicated for the cluster visualization interface, user query processing mechanisms, and an integrated AI chatbot powered by a compact, fine-tuned LLM.

Processing the full archive entails analyzing over 14,000 pages exhibiting considerable non-uniformity in structural organization, layout, and writing styles. This variability complicates embedding generation and necessitates a collaboration between humanities researchers possessing familiar with the archive and the topics it covers, and technical developers responsible for the implementation of the system.

Future iterations should address the system's current limitations through prototype learning mechanisms, providing the ability to reduce algorithmic bias, enhance classification accuracy, and enable scholars to manually correct misclassifications, thereby improving system reliability. While the architecture permits adaptation to comparable censorship archives, such migration would require a complete recomputation of document and page embeddings,

constituting a full analysis with minimal elements of reuse.

FINAL REFLECTION

The project goal was to make a mechanism of censorship in Francoist radio both visible and measurable, and, in this sense, it successfully met its initial objectives. Why and how instead of where and when. Due to a lack of time, we only based our model on the year 1942.

By combining archival research with AI-based visual and semantic analysis, the work demonstrates that censorship can be studied not only through isolated historical interpretation but also through systematic, data-driven observation. The resulting interactive map offers a new way of approaching historical documents, transforming dispersed and fragile materials into an accessible structure that reveals patterns of suppression, thematic priorities, and ideological control. On a personal and critical level, the project highlights how technology can serve memory rather than erase it, allowing silenced voices and hidden interventions to emerge in contemporary public understanding.

At the same time, the project remains conditioned by important limitations. The absence of OCR restricts direct textual interpretation, meaning that classifications depend heavily on visual and contextual cues rather than precise linguistic analysis. The manual definition of thematic clusters and heuristic estimation of censorship introduce interpretative bias, and the dataset itself cannot fully represent the diversity of radio production under the dictatorship. These constraints open further questions: how might results change with larger or more diverse archives, with multilingual textual analysis, or with collaborative historical validation? To what extent can AI truly interpret silence, absence, or intention in censored discourse?

Overall, the project should be understood not as a definitive answer to the history of censorship, but as a methodological proposal. Its main contribution lies in demonstrating that digital humanities tools can expand historical inquiry, encouraging future research that deepens, questions, and refines the relationship between technology, archives, and democratic memory.