

Digital Archeology: LLM-Driven Relational Modeling of Banking Archives

Andrea Colombo¹, Pietro Pinoli¹

¹Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Via Giuseppe Ponzio 34, 20133 Milan, Italy

Abstract

Historical banking archives contain rich evidence for the study of financial institutions, governance, and political economy, yet their analysis remains severely constrained by heterogeneous document formats, degraded scans, and evolving institutional structures. We present a pragmatic, end-to-end conceptual modeling and data engineering effort that transforms 182,719 scanned pages of Italian banking supervision records from 1910–1960 into a structured, document-centric relational database. The effort combines Vision Language Models and Large Language Models in a six-stage *sequential* VLM→LLM pipeline that covers OCR, document segmentation, metadata classification, data integration, governance extraction, and credit exposure extraction. Its conceptual model comprises a central document entity linked through explicit foreign keys to governance entities and credit exposures, balancing expressive representation of historical relationships with the practical constraints of automated LLM-based extraction. The resulting database contains more than 430,000 records spanning 778 banking institutions, including 34,771 governance entities and 212,760 credit exposures.

Keywords

Conceptual Modeling, Relational Model, Large Language Models, Data Extraction Pipeline, Historical Archives

1. Introduction

The archives of the Italian Banking Supervision Authority from Banca d'Italia, the Central Bank of Italy, span seven decades (1910–1960) [1] and contain inspection reports, board meeting minutes, financial statements, loan portfolios, and correspondence between banks and the Bank of Italy, resulting in a set of over 182k scanned images. These materials are of immense value to economic historians, banking regulators, and political scientists studying regulatory capture and the political economy of banking. Yet they remain largely inaccessible for systematic analysis: on one side, the documents exist only as scanned JPEG images, requiring manual review of degraded historical typography. On the other, only a conceptual model and a structured database can allow us to conduct overall analysis on the data.

This paper presents a (i) conceptual modeling and (ii) data engineering effort that bridges this gap. To this aim, inspired by similar works [2], we design a pragmatic six-stage sequential VLM→LLM pipeline that transforms raw scanned images into a structured relational database of the banking supervision documents [3]. The resulting database comprises over 430k records across three core entities: `documenti` (documents), `organismi_sociali` (governance bodies), and `fidi_prestiti` (credit exposures). The engineering contribution of this work to the conceptual modeling community is threefold:

1. We propose a **document-centric conceptual schema** for heterogeneous archival corpora that balances expressiveness, analytical utility, and practical constraints of LLM-based extraction, linking knowledge through a denormalized design with lightweight intra-archive name harmonization and filename-anchored joins.
2. We identify and formalize **conceptual modeling challenges** specific to historical banking archives—including temporal schema evolution across 50 years of institutional discontinuities, domain constraints, and entity normalization under noisy OCR conditions.
3. We release the complete pipeline code on GitHub¹ and provide analytical queries for potential use cases.

ER2026: Companion Proceedings of the 45th International Conference on Conceptual Modeling: Doctoral Consortium, Tutorials, Posters and Demos, October 5-8, 2026, St. John's, NL, Canada

✉ andrea.colombo1@polimi.it (A. Colombo); pietro.pinoli@polimi.it (P. Pinoli)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://github.com/andreac0/ispezioni-banche-storiche> (documentation in English; archival source texts remain in Italian)

2. Background and Motivation

The source archive contains thousands of digitized inspection reports, correspondence, and financial documents from the Banca d’Italia’s supervisory archives. These cover a heterogeneous banking landscape: commercial banks (Banco di Napoli, Credito Italiano, Banco Lariano), cooperative credit banks (*Casse di Risparmio*, e.g., Cassa di Risparmio di Pescara), Monti di Credito su Pegno, and agricultural credit institutions. The period encompasses the Fascist era, World War II, and the post-war reconstruction—each introducing distinct challenges for metadata extraction (e.g., Fascist-era calendar dates such as “Anno XV” requiring conversion to Gregorian years).

The core conceptual modeling challenge lies in transforming heterogeneous, noisy OCR text into a consistent relational schema. This process must navigate several domain-specific hurdles, beginning with lightweight name harmonization (which is different from a full entity resolution), where bank names appear with inconsistent spellings across documents from the same archive unit (e.g., “Credito Italiano” versus “Credito Italiano-Vigilanza”). Furthermore, key connections such as board memberships, family ties, and political affiliations are embedded entirely within free text rather than structured fields. The schema must also strictly enforce domain constraints; for instance, insider lending should only be flagged when a debtor serves as a board member of the *same* bank that granted the credit, rather than an external entity. Finally, because the source documents span 50 years of evolving institutional structures, the underlying model must accommodate these historical discontinuities while preserving longitudinal data integrity.

3. Conceptual Schema Design and Pipeline Architecture

The archival corpus is heterogeneous by nature: inspection reports, correspondence, financial statements, and loan portfolios are produced by a diverse set of persons, span five decades, and follow evolving document conventions or even inspector reporting preferences. A viable conceptual model must therefore satisfy several competing requirements.

- **Expressiveness:** the schema must capture both the obvious document-level metadata (bank, date, type) and the rich relational structure hidden inside free-text inspection reports—board compositions, credit exposures, political affiliations, and family ties.
- **Analytical utility:** the most valuable research questions in economic history and political economy involve cross-entity queries, so, for instance, “which banks had board members with Fascist party affiliations, and what was the credit quality of loans granted to those same individuals?” The schema must make such joins natural and efficient.
- **Practical constraints:** the schema must be built by considering that an automated LLM-based pipeline will have to populate it with acceptable error rates, which leads us to build a denormalized design with fewer tables and explicit foreign keys over a deeply normalized schema.

Balancing these requirements, we adopt a *document-centric* design in which a central domain concept anchors all extracted information, with satellite concepts for relational content discovered by LLM extraction [4].

Conceptual (domain) model. Domain concepts are *Bank*, *Document*, *Person*, *Role*, *BoardMembership* (with validity interval), *CreditExposure* and *Guarantee*. Relations: *Bank–Document*, *Person–Role–BoardMembership–Bank*, *Person–CreditExposure*, family ties as recursive *Person* association, political affiliation as *Person* attribute. Evolving bank types and fascist calendar dates (“Anno XV”) are modeled as attributes, and historical discontinuities are preserved by anchoring provenance to *Document*.

Logical implementation. Figure 1 shows the *logical* denormalization of this conceptual model for LLM population: `document_i` anchors provenance and preserves longitudinal integrity across 50 years of institutional discontinuities (Sec. 2), while `organi_sociali` materializes *BoardMembership* and `fidi_prestiti` materializes *CreditExposure*; *Bank* is not a separate table to avoid unstable identity. Both satellite tables link via FK `filename`.

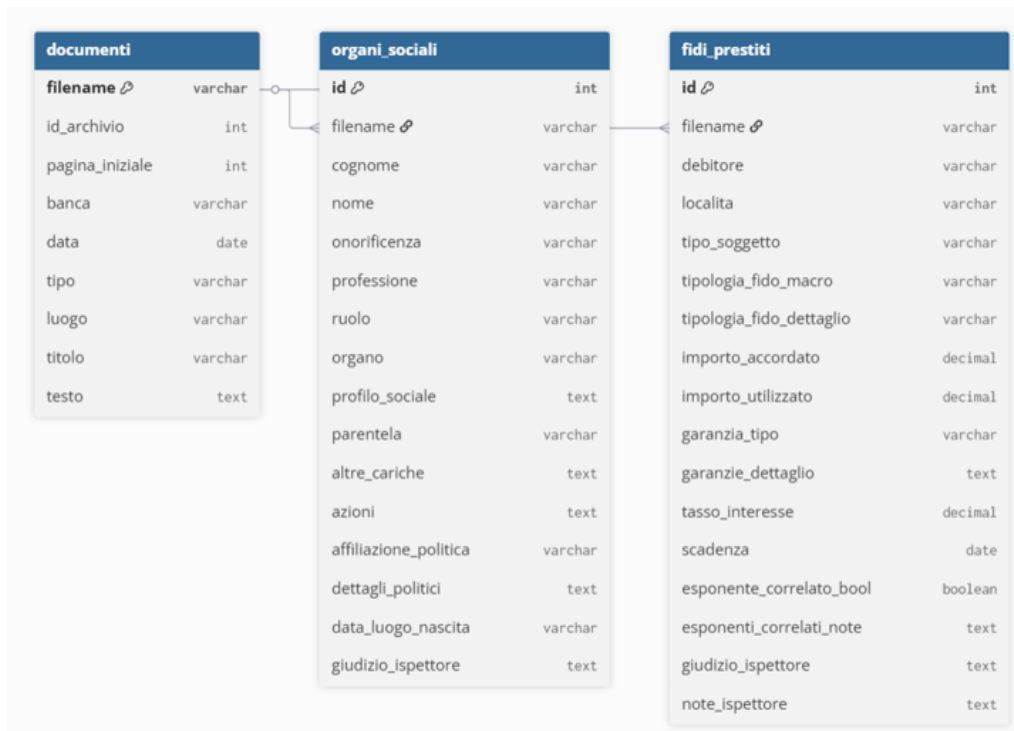


Figure 1: Logical relational schema (denormalization of the domain model). document i is central; governance and credit exposures link via FK filename. Fig. is at the logical layer; see text for the conceptual model.

Core entity: document i - Each row is one archival document (inspection report, letter, financial statement). Attributes: harmonized bank name, date, type, synthetic title, full OCR text. PK is filename, the single source of truth for provenance enabling temporal/geographic analysis.

Governance entity: organi_sociali. Each row is one person (board member, supervisor, senior manager). Attributes: name, role (*President, Advisor...*), honorifics, titles, political affiliations, family ties, inspector evaluations. FK filename enables joins such as “who served on multiple banks?”.

Credit exposure entity: fidi_prestiti. Each row is one loan. Attributes: debtor, loan type, amount, guarantees, rate, inspector quality (regular/doubtful/NPL/loss). FK filename links to source document; insider lending is detected by joining fidi_prestiti debtor name against organi_sociali names for the same filename (same-bank constraint).

3.1. Pipeline Architecture

With the target schema defined, the pipeline can be designed to produce structured outputs that map directly onto the three tables. The pipeline follows a batch-oriented ETL pattern with six sequential stages, each producing structured outputs that feed into downstream processes (Figure 2).

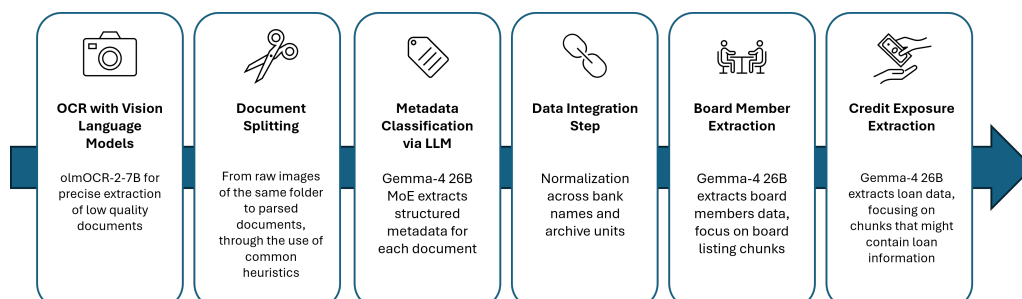


Figure 2: Sequential stages to transform scanned JPEG images to a structured relational database.

Stage 1: OCR with Vision Language Model Scanned document images (JPEG) are processed by

the o1mOCR-2-7B model [5] via vLLM. Images are resized to a maximum dimension of 1,288 px and processed in batches of 128. The model handles historical Italian typography, degraded scans, and complex document layouts. Output: flat text files per archive folder, with page-level markers.

Stage 2: Document Splitting Raw OCR output is parsed and split into individual documents/letters using heuristic rules. The splitter identifies document boundaries through formal Italian closing patterns (“*distinti ossequi*”, “*cordiali saluti*”) and opening markers (“*OGGETTO*”).

Stage 3: Metadata Classification Each text fragment is classified by Google’s Gemma-4 LLM (26B MoE, 4B active parameters), which extracts structured metadata: bank name, date, document type, location, and a synthetic title.

Stage 4: Data Integration Metadata is merged with the original OCR text content. Bank names are harmonized across documents from the same archive unit using **frequency-based normalization**: within each archive prefix, the most frequent name is selected as canonical, discarding OCR variants.

Stage 5: Board Members & Governance Extraction The Gemma-4 model analyzes inspection reports to extract board members (*CdA*), supervisory board, and senior management. A smart chunking strategy focuses on relevant text segments using keyword detection (“consiglio”, “amministrazione”, “presidente” ...), processing chunks of 12,000 tokens with 2,000-token overlap. Strict exclusion rules filter out customers, debtors, and Bank of Italy personnel.

Stage 6: Credit Exposure Extraction The LLM extracts detailed loan and credit information: debtor identities, loan types, amounts, guarantees, interest rates, and inspector quality assessments (regular, doubtful, NPL, loss). A critical feature is the identification of **insider lending**—credit granted to board members or their families.

Design decisions, tradeoffs and lessons. Batch reproducibility on one A100 80 GB (vLLM 0.19.0) drove choices: o1mOCR-2-7B local (batch 128, 1288 px), Gemma-4 26B-MoE/4B-active (greedy, 32k ctx, prefix caching, chunked prefill) for determinism and long reports; heuristics (closing patterns “*distinti ossequi*”, 20% unique-word filter) handle degraded scans and hallucination loops. All inference uses greedy decoding (temp. 0.0); 8,192/32,768 max tokens for OCR/extraction.

4. Key Findings

Findings below are *descriptive counts from LLM extraction* without formal precision/recall evaluation. The initial analysis reveals patterns in the banking system, summarized in Table 1.

Scale and Diversity. The dataset captures **778 unique banking institutions**, from major national banks to small local cooperative banks. Banco di Napoli leads with 152 inspection reports, followed by Credito Italiano and Banca Nazionale del Lavoro (98 each), and Banco Lariano (83).

Credit Quality. Across the dataset, **35,769 credit exposures** are classified as non-performing or definite losses. The Cassa di Risparmio di Pescara e di Loreto Aprutino tops the list with 430 critical exposures and a combined exposure value exceeding 3.3 billion historical lire.

Insider Lending. A striking **27,542 credit exposures** are flagged as *potential* insider lending (same-filename name join, indicative)—loans to board members, particularly in smaller banks.

Political Entanglements. The extraction indicates deep Fascist penetration: **673** individuals as “Regime Fascista”, **121** as Antifascist, and **48** as explicit PNF members (as extracted).

Table 1

Key findings from the extracted dataset.

(a) Dataset scope		(b) Governance & politics		(c) Credit quality	
Metric	Value	Metric	Value	Metric	Value
Unique banks	778	Insider lending	27,542	Critical exposures	35,769
Inspection reports	9,705	Multi-bank directors	617	Total value (hist. lire)	~858 bn
Pages processed	182,719	Fascist-linked	721	Mortgage criticality	28.6%
Governance entities	34,771	Antifascist	121	Unsecured failure	12.2%
Credit exposures	212,760				

The pipeline has been run end-to-end on the full corpus (430k+ records). Code and 15 queries are on GitHub as six Python scripts with reproducible GPU config. The full dataset is available upon request to the Banca d'Italia Historical Archives. Preliminary manual inspection of over 30 documents showed high accuracy for structured fields (bank names, dates, roles).

5. Conclusion and Future Work

We have presented a pragmatic relational pipeline that transforms scanned Italian banking supervision images into a structured database. The contribution is primarily an engineering and conceptual-modeling effort: a document-centric schema balancing expressiveness with LLM-population constraints, implemented as a sequential VLM→LLM pipeline, capable of addressing the unique challenges of historical archival data.

In future works, we plan to improve the corpus generation by (i) conducting formal precision/recall evaluation; (ii) generating interactive network visualization of family ties, multi-bank directorships, and insider lending; (iii) conducting temporal analysis of credit quality evolution; (iv) exposure via a REST API for integration with external research tools; (v) export in standard formats (CSV, JSON, RDF) for interoperability with academic data repositories; and (vi) extension to further additional archive folders [6].

Acknowledgments

We thank the Banca d'Italia Historical Archives for access to the scanned documents. This research was supported by the project BETTER (Grant agreement n. 101136262).

Declaration on Generative AI

During the preparation of this work, the authors used LLM-based tools for grammar and spelling checks. The authors reviewed and edited the content as needed and take full responsibility for the content.

References

- [1] M. Molteni, D. Pellegrino, The establishment of banking supervision in Italy: an assessment (1926–1936), *Business History* 66 (2022) 1442–1470. URL: <https://ora.ox.ac.uk/objects/uuid:fe315c02-ec90-4bbb-a5df-5642e1907046>. doi:10.1080/00076791.2022.2134347.
- [2] M. M. Nishat, P. S. Nervik, N. O. Olsson, B. S. Andersen, A. Rauzy, A semantically structured relational database for construction safety analysis, *Construction Innovation* (2026) 1–33. URL: <http://dx.doi.org/10.1108/CI-06-2025-0259>. doi:10.1108/ci-06-2025-0259.
- [3] P. Guimarães, A. C. Vieira, V. Sousa, G. Moreira, M. Y. Santos, Automated and llm-assisted conceptual modeling from semi-structured data, in: *IFIP Working Conference on The Practice of Enterprise Modeling*, Springer, 2025, pp. 121–129.
- [4] J. Silva, Q. Ma, J. Cabot, P. Kelsen, H. A. Proper, Towards Human-in-the-Loop LLM-Enabled Domain Modeling, Springer Nature Switzerland, 2025, p. 127–145. URL: http://dx.doi.org/10.1007/978-3-032-08623-5_7. doi:10.1007/978-3-032-08623-5_7.
- [5] Allen Institute for AI, olmOCR-2-7B: A vision language model for document OCR, <https://huggingface.co/allenai/olmOCR-2-7B-1025>, 2025. Accessed: 2026-06-15.
- [6] Banca d'Italia, The historical archives, 2026. URL: <https://www.bancaditalia.it/chi-siamo/asbi/index.html?com.dotmarketing.htmlpage.language=1>, accessed 2026-07-22.