The Cartoonists Meta-description Framework

Valerio De Luce[1,†], Tiberio Uricchio[2,†]

[1]Università di Macerata, Via Crescimbeni, 30/32 - 62100 Macerata, Italia
[2]Università di Pisa, Largo Lazzarino, 1 - 56122, Italia

Abstract

We introduce the Cartoonists Meta-description Framework (CMF), an agentic system that automates the creation of standards-compliant, hierarchically aware metadata for cartoonists' archives. These archives pose unique challenges: their materials, which may consist of sketches, drafts, annotated agendas, finished cartoons, contracts, reflect creative rather than bureaucratic processes, and their meaning emerges from the interplay of image, text, and narrative sequence. The CMF coordinates multiple Large Vision Language Models (LVLMs) to generate structured descriptions, extract named entities, and classify document types. We evaluate the framework on the Sergio Staino Archive "Satira e Sogni", processing 300 digitized images across two experimental configurations. The results reveal a trade off: multi agent architectures with large models achieve better entity extraction, while an ensemble of smaller specialized models produces descriptions with higher terminological alignment. Entity extraction remains challenging, particularly for indirect references that require cultural and historical knowledge. These findings position the CMF as an augmentation tool that accelerates archival workflows while preserving the central role of professional expertise.

1. Introduction

Describing cartoonists' archives and collections is a complex endeavor that current AI systems only partially address. Unlike traditional archives, the collections of cartoonists reflect creative processes [1], and their "Fonds Creator" tend to accumulate a wide variety of heterogeneous materials: sketches, drafts, annotated diaries and agendas, finished cartoons, but also more orthodox material such as contracts and correspondence. The meaning of these archives emerges from the interplay of image, text, layout, and narrative sequence, corresponding to a series of multi-modal challenges that purely textual approaches cannot address. Institutions face trade offs between access and detailed description [2]. Similar tensions characterize artists' archives more broadly [3], where digitization creates both opportunities and demands for sophisticated descriptive strategies.

These reasons have led recent AI applications in archival contexts to focus predominantly on textual documents, using natural language processing for classification [4] and metadata extraction [5].

Research on computational comics understanding has similarly targeted specific tasks such as panel segmentation, text extraction, and layout analysis [6]. More broadly, Large Vision-Language Models (LVLMs) offer promising capabilities for visual analysis; however, their direct application to archival description risks hallucinated content, terminological inconsistencies, and structurally inadequate metadata [7].

We propose the Cartoonists Meta-description Framework (CMF), an agentic system that coordinates multiple LVLMs to generate standards-compliant metadata. To our knowledge, this is the first study that applies agentic systems to the archival description of cartoonist's materials using multiple LVLMs. We evaluated the CMF on the Sergio Staino Archive "Satira e Sogni", testing 300 digitized images across two experimental configurations. Our results provide evidence of both capabilities and limitations of current zero-shot approaches, with implications for other visual and artistic archives facing similar challenges.

IRCDL 2026: 22nd Conference on Information and Research Science Connecting to Digital and Library Science, February 19-20, 2026, Modena, Italy

v.deluce@unimc.it (V. D. Luce); tiberio.uricchio@unipi.it (T. Uricchio)
0009-0006-6190-6406 (V. D. Luce); 0000-0003-1025-4541 (T. Uricchio)

2. Related Work

The application of AI to archival metadata is an emerging field. Groppe et al. [8] proposed a federated multi agent framework that coordinates large language models to generate ISAD(G)-compliant metadata from textual inventories. Although innovative, this approach addresses homogeneous textual materials and lacks visual perception capabilities, making it ill-suited to the graphic complexity of cartoonists' archives. The broader implications of AI adoption in the archival field have been examined by InterPARES Trust AI [9], which investigates questions of archival authenticity, reliability, and trust. Other studies have demonstrated that LLMs' have great potential in automated keyword generation and metadata enrichment in museum collections [10]. Such approaches consistently emphasize the fact that automation should accelerate workflows while keeping archival expertise central, a position aligned with the Augmented Humanities paradigm [11], which frames AI as collaborator rather than replacement.

Vision Language Models have transformed computational approaches to visual archives. Models such as CLIP [12], BLIP [13] and LLaVA [14] enable captioning, visual question answering, and semantic enrichment of digitized images. Recent work has applied LVLMs to interpretable search in visual cultural heritage collections [15], and these capabilities have also been tested in museums and digital libraries [16]. However, heritage contexts require strict terminological consistency [17], and current models frequently violate controlled vocabularies or hallucinate content. Integrating archival tools such as thesauri and ontologies with generative models remains an open challenge.

Research on computational comics understanding has advanced considerably, driven by the medium's complexity as an AI testbed [18]. Work has addressed panel segmentation, text extraction, and reading order detection [19], while methodologies such as "distant viewing" [20] have expanded analytical possibilities. However, these methods target low-level perception rather than archival requirements such as provenance contextualization and adherence to archivist professional standards.

3. The Cartoonists Meta-description Framework

The CMF coordinates multiple LVLMs through specialized agents that collaborate to produce item level descriptions. The system operates in zero-shot mode with a fixed set of on premises models, chosen to ensure data sensitivity: copyright and institutional agreements prohibit transmission to external cloud services, a common restriction for cultural heritage collections.

3.1. Source Materials

The input to the CMF consists of high-resolution scans of physical archival materials. These are not generic digital images but reproductions of paper documents with specific material properties that archival description must capture. Cartoonists' archives typically contain heterogeneous document types: sketches and drafts (preliminary drawings on notebook pages, loose sheets, or cardboard, often with annotations and corrections), finished vignettes (completed single-panel cartoons in ink, sometimes with wash or color), comic strips (multi-panel sequences), and preparatory materials (reference photographs, clippings, traced layouts).

3.2. The Architecture

The framework comprises five agents (Figure 1). The Orchestrator coordinates the overall workflow and distributes tasks among the other components. The DescriptionAgent is responsible for generating archival descriptions according to the reference schema, while the EntityAgent extracts named entities in alignment with the controlled vocabulary. The ClassificationAgent determines the typology of each document, and the ValidatorAgent ensures that the outputs are complete and consistent. Together, these agents form an integrated system that supports structured, reliable, semantically rich process in compliance with archival principles applied to digital records. The CMF starts with an input consisting

of image files and a prompt containing the operating instructions for compiling the descriptive schema. This schema is composed of the descriptive fields present in the original archive and required for creating description sheets, which are: date (*estremo remoto*), type, medium, description, height, width, and document type. We also request that the following entities be extracted: organizations, people, notable things, events, documents, aggregations, and roles. Finally, we request the types of relationships, which include: people, organizations, places, events, families, and notable things.

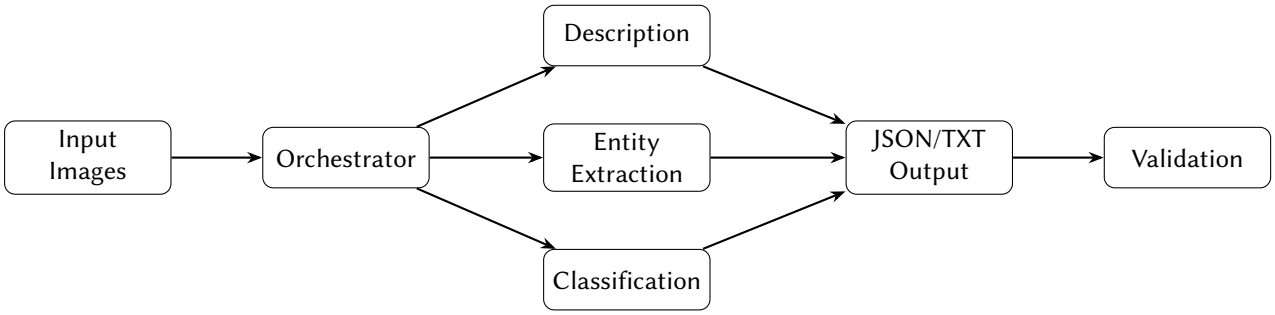

Figure 1: CMF framework. The Orchestrator distributes tasks to the agents; outputs are aggregated into JSON metadata and then validated.

Each model in the two configurations we tested receives images with structured prompts that include the controlled vocabulary and require specific output fields. Since fewer than five models are used in each configuration, a single model may fulfill more than one task; we nonetheless refer to each functional role as an agent. The Entity Extraction agent identifies entities with their types and relationship categories: *depicted* (visually represented), *mentioned* (named in text within the image), or *referred* (indirectly referenced through allusion or satirical commentary). This three-way classification reflects standard practice for documenting entity-record relationships, with “referred” representing the most interpretive and challenging category. The system treats the existing archival hierarchy as authoritative, using it as contextual framework rather than attempting to discover structure.

4. Case Study: The Sergio Staino Archive

The Sergio Staino Archive "Satira e Sogni" provided an ideal case study: a carefully curated collection combining a traditional hierarchical structure with rich item level descriptions, including typed entity relationships.

From the "Fondo archivio cartaceo", we collected 300 digitized images spanning sketches, vignettes, and strips, organized to mirror the original archival order.

The sample was stratified by document type (54% sketches, 15% vignettes, 18% strips, 13% other) across 101 archival units.

From the existing metadata, we derived a controlled vocabulary of 27 entities: 14 persons (including Staino’s recurring characters Bobo, Bibi, Michele, Ilaria), 7 organizations, 3 places, 2 objects, and 1 event. For each of the 300 images it contains the original fields, including entities and relationship types.

Ground truth entity-unit associations were classified according to three relationship types that reflect how entities connect to archival records: i) *depicted*: the entity is visually represented in the image (e.g., a drawing of Bobo, a recognizable caricature of a politician); ii) *mentioned*: the entity’s name appears explicitly in text within the image, such as dialogue balloons, captions or handwritten annotations; iii) *referred*: the image alludes to the entity through indirect reference, satirical commentary, or contextual implication without explicit visual or textual identification (e.g., a cartoon about political scandal may refer to the politician involved without depicting or naming them).

The distribution across these types is referred (42%), depicted (30%), and mentioned (28%). This reflects the heavily allusive nature of Staino’s satirical work and presents a significant challenge for automated systems, as “referred” relationships require cultural and historical knowledge.

5. Experimental Setup

5.1. Hardware and Models

All experiments were conducted on a workstation-level machine (NVIDIA DGX-Spark with 128 GB of memory) rather than datacenter infrastructure, reflecting resources accessible to small and medium-sized cultural institutions. We selected several state of the art open source models, in order to run them on our hardware. This choice is driven by the need to maintain privacy of the images, as required in many archive contexts. We chose as primary model Qwen2.5-VL-72B [21], 4-bit quantized to fit in our machine. Temperature was set to 0.1 across all agents to minimize hallucinations.
We also evaluated Qwen2.5-VL-7B, LLaVA-13B [22], Llama-3.2-Vision-11B [23], and MiniCPM-V-8B [24]. Model assignment was determined empirically based on output quality and reliability. The Orchestrator coordinates agents without invoking an LLM. Llama-3.2-Vision-11B handles validation in both configurations.
We chose zero-shot evaluation to assess baseline capabilities without preliminary annotation, precisely the labor the system aims to reduce. Recent work on zero-shot NER [25] suggests self-improving frameworks can enhance performance without fine-tuning, though such approaches remain untested in archival and graphic cultural heritage contexts.

5.2. Test Configurations

We conducted five experimental runs in two configurations (Table 1). Runs 1 to 4 employed identical model configurations to assess output variability and stability across repeated executions. Then, the run number 5 tested an ensemble of smaller specialized models to determine whether task specialization could compensate for lower model scale. The Orchestrator is hard-coded in all our experiments.

Config.	Task	Model
Runs 1-4	Description	Qwen2.5-VL-72B
	Entity Extraction	Qwen2.5-VL-72B
	Classification	LLaVA-13B
	Validation	Llama-3.2-Vision-11B
Run 5	Description	Qwen2.5-VL-7B
	Entity Extraction	MiniCPM-V-8B
	Classification	LLaVA-7B-v1.6
	Validation	Llama-3.2-Vision-11B

Table 1
Task-to-model assignment.

5.3. Evaluation Protocol

Quantitative metrics were calculated against the ground truth and controlled vocabulary using commonly used classification and textual metrics. For the entity extraction task, it was evaluated using standard named entity recognition metrics computed at the document level: precision, recall, and F1-score. Description quality was assessed using two semantic similarity measures that capture different aspects of textual alignment: BERTScore [26] and Sentence-BERT [27]. BERTScore rewards terminological precision, while Sentence-BERT rewards semantic equivalence. We also conducted a human evaluation to manually verify the descriptions. In this specific case, a Likert scale was used: 1 = inadequate, 2 = partially adequate, 3 = adequate.

6. Results

6.1. Entity Extraction

Table 2 reports NER performance for all the tests. Run 4 achieved the best F1 score (20.7%), with the highest precision among multi agent configurations. Run 5 showed higher recall (26.7%) but lower precision (9.9%), producing 616 raw extractions of which only 164 matched the controlled vocabulary. Precision is calculated at document level: an extraction is correct only if both the entity and its assignment to the specific document match ground truth. This explains Run 5's trade off as it captures more entities overall but with many incorrect document assignments.

Configuration	Precision	Recall	F1
Run 4 (multi agent, best)	21.4%	20.0%	20.7%
Runs 1–4 (avg.)	*17.0%*	*20.0%*	*18.0%*
Run 5 (Ensemble)	9.9%	26.7%	14.4%

Table 2
Entity extraction performance. Runs 1-4 use identical settings; F1 ranges from 15.4% to 20.7% ($\pm$2.1 pp).

Error analysis reveals systematic patterns. By entity type, persons achieved the highest recall (Bobo, Bibi, Michele, all visually distinctive recurring characters), while organizations and events were rarely identified. By relationship type, "depicted" entities (visually present) achieved F1 $\approx$ 12%, "mentioned" (text in balloons) F1 $\approx$ 5% and "referred" (indirect references requiring cultural knowledge) F1 $\approx$ 0%. The complete failure on "referred" relationships, which constitute 42% of ground truth associations, explains much of the overall low performance. False positives frequently included generic descriptors ("uomo", "donna") or hallucinated entities not present in the controlled vocabulary.
The most recognized entities were Staino's recurring characters: Bobo (67 extractions in Run 5), Michele (19), Ilaria (21). However, these were frequently over-extracted across documents where they did not appear. Historical figures (Berlinguer, Paolo VI) appeared consistently in false negatives, reflecting the models' inability to recognize Italian political caricatures without cultural context.

6.2. Description Quality

Table 3 reports semantic similarity metrics across all the test. An inverse trade off emerged between the two configurations: the ensemble (Run 5) achieved substantially higher BERTScore F1 (71.6% vs. 34.5% average for Runs 1 to 4), indicating better token level alignment with professional archival descriptions. Conversely, the multi agent configurations achieved higher Sentence-BERT similarity (72.1% best, 71.3% average vs. 65.8%), capturing overall semantic content more effectively.

Configuration	BERT-P	BERT-R	BERT-F1	Sent-BERT
Run 4 (multi agent, best)	34.6%	36.7%	35.6%	72.1%
Runs 1–4 (avg.)	*33.7%*	*35.5%*	*34.5%*	*71.3%*
Run 5 (Ensemble)	71.0%	72.1%	71.6%	65.8%

Table 3
Description quality metrics.

The divergence between metrics reflects different generation strategies. Smaller specialized models produce shorter, more formulaic descriptions that closely replicate archival terminology (higher BERTScore, lower semantic richness). Larger models generate more varied descriptions that paraphrase rather than replicate reference terminology (lower BERTScore, higher Sentence-BERT). This suggests a practical trade off: Run 5 outputs require less terminological revision but may need semantic enrichment, while Runs 1-4 outputs capture meaning well but require terminology standardization.

6.3. Human Evaluation

Human evaluation confirmed patterns observed in quantitative metrics (Table 4). Run 5 showed improved description quality (2.7/3, 85% adequacy) but weaker entity recognition (1.6/3, 30%). Document classification showed marginal decline between configurations.

Task	Run 4		Run 5	
	Score	Adequacy	Score	Adequacy
Description Generation	2.4	70%	2.7	85%
Document Classification	2.6	80%	2.5	75%
Entity Recognition	1.8	40%	1.6	30%

Table 4
Human evaluation results on Runs 4 and 5.

The system reliably extracted information directly observable in images or derivable from file paths: dates, document type, support, position in the aggregate, and visual content. It consistently failed on physical dimensions (requiring direct measurement unavailable from scans), entity identification (particularly indirect references), and relationship type classification (with performance degrading from visually evident to culturally embedded references).

7. Discussion

7.1. Configuration Trade offs

Our results reveal a consistent trade off: model scale enhances entity extraction precision, while task specialization with smaller models improves terminological alignment. This finding has practical implications. Workflows prioritizing comprehensive entity capture may favor the ensemble approach (higher recall), accepting the cost of human filtering. Workflows prioritizing draft quality may favor larger models (higher Sentence-BERT), accepting weaker entity extraction.

The complete failure on "referred" relationships deserves attention. Staino's satirical work is dense with indirect references to Italian political figures and period-specific events. Recognizing that a caricature "refers to" Craxi's scandals or Berlinguer's *compromesso storico* requires historical and cultural knowledge that visual analysis cannot provide. This limitation is fundamental rather than technical: zero-shot approaches lack the contextual grounding that archival description requires.

7.2. Practical Value

The CMF's contribution lies not in replacing professional judgment but in accelerating routine workflow phases. The two configurations serve complementary purposes: the multi-agent pipeline, with higher Sentence-BERT similarity (72.1%), produces descriptions that better capture overall semantic content, providing a more complete draft for expert refinement; the ensemble, with higher BERTScore F1 (71.6%), generates output more closely aligned with the lexical patterns of existing archival descriptions, facilitating direct integration into institutional metadata. Pre-classifying document typology at 87% accuracy further reduces manual effort. The ensemble's over-extraction behavior: high recall, low precision, may suit workflows where human review is expected, capturing potentially relevant entities for validation rather than risking omission.

These findings underscore that current zero-shot approaches cannot substitute archival expertise. The archivist brings irreplaceable knowledge: the creator's working methods, period-specific references, political caricature conventions, and terminological consistency with institutional standards. The CMF functions as an augmentation tool that handles routine tasks and ensures structural compliance while depending on human validation for contextual accuracy.

8. Conclusions

We presented the Cartoonists Meta-description Framework and evaluated it on the Sergio Staino Archive. Multi agent architectures with large models achieve better entity extraction; ensembles of smaller specialized models produce higher terminological alignment. Entity extraction in zero-shot settings remains limited, particularly for indirect references requiring cultural knowledge. The main finding is methodological: cartoonists' archives, deeply embedded in specific cultural contexts, resist purely computational description.

Consistent with the Augmented Humanities paradigm, the CMF functions as an augmentation tool, generating preliminary descriptions and ensuring structurally compliant output while depending on human expertise to ensure contextual validity.

A promising direction is the integration of Retrieval-Augmented Generation (RAG) techniques [28], which could provide agents with external knowledge sources, biographical profiles, historical context, archival thesauri, to ground entity extraction in factual evidence rather than relying solely on visual inference.

9. Acknowledgments

Work partially supported by the Italian Ministry of Education and Research (MUR) in the framework of the FoReLab project (Departments of Excellence). We gratefully acknowledge NVIDIA Corporation for providing the DGX Spark system used in this research.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] J. Douglas, Archiving creative practice: notes toward a research agenda for historians, Archivaria 76 (2013) 127–152.

[2] M. A. Greene, D. Meissner, More product, less process: Revamping traditional archival processing, The American Archivist 68 (2005) 208–263.

[3] C. Damiani, Gli archivi dell'arte: Gestione e rappresentazione tra analogico e digitale, volume 1, Editrice Bibliografica, 2023.

[4] T. Hutchinson, Natural language processing and machine learning as practical tools in the archival processing of email, The American Archivist 83 (2020) 7–29.

[5] G. Büttner, Auto-classification in an international organization: report from a feasibility study, Comma 2017 (2019) 15–26.

[6] A. Dutta, S. Biswas, A. K. Das, Cnn-based segmentation of speech balloons and narrative text boxes from comic book page images, International Journal on Document Analysis and Recognition (IJDAR) 24 (2021) 49–62.

[7] S. Allegrezza, Intelligenza artificiale e archivi: opportunità e rischi, JLIS.it 16 (2025).

[8] S. Groppe, et al., Automated archival description with federated large language models, Journal of Documentation (2025).

[9] P. Feliciati, et al., Interpares trust ai: exploring artificial intelligence for archival authenticity, Archival Science 21 (2021) 389–408.

[10] D. van Strien, et al., Large language models to make museum archive collections more accessible, AI & Society (2025).

[11] R. Bryant, et al., Augmented humanities: Human-machine collaboration in the archive, Digital Humanities Quarterly 15 (2021).
[12] A. Radford, J. W. Kim, C. Hallacy, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, 2021, pp. 8748–8763.
[13] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International Conference on Machine Learning, 2022, pp. 12888–12900.
[14] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, 2023. URL: https://arxiv.org/abs/2304.08485. arXiv:2304.08485.
[15] B. nationale de France, et al., Explainable search and discovery of visual cultural heritage collections with multimodal large language models, arXiv preprint arXiv:2411.04663 (2024).
[16] M. Wevers, T. Smits, The visual digital turn: Using neural networks to study historical images, Digital Scholarship in the Humanities 35 (2020) 194–207.
[17] P. Harpring, Introduction to controlled vocabularies: terminology for art, architecture, and other cultural works, Getty Publications, 2013.
[18] E. Vivoli, et al., One model to rule them all: Towards universal visual understanding of comics, Pattern Recognition (2024).
[19] Y. Zhang, S. Hotta, Automatic reading order detection of comic panels, in: International Conference on Pattern Recognition, Springer, 2022, pp. 76–90.
[20] T. Arnold, L. Tilton, Distant viewing: Analyzing large visual corpora, Digital Scholarship in the Humanities 34 (2019) i28–i40.
[21] Qwen Team, Qwen2.5-vl technical report, arXiv preprint arXiv:2502.13923 (2025).
[22] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, Advances in Neural Information Processing Systems 36 (2023).
[23] Meta AI, Llama 3.2: Revolutionizing edge ai and vision with open, customizable models, https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/, 2024.
[24] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He, et al., Minicpm-v: A gpt-4v level mllm on your phone, arXiv preprint arXiv:2408.01800 (2024).
[25] T. Xie, Q. Li, Y. Yan, J. Liu, H. Wang, Self-improving for zero-shot named entity recognition with large language models, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2024, pp. 583–595.
[26] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675 (2019).
[27] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of EMNLP-IJCNLP, 2019, pp. 3982–3992.
[28] G. Di Marcantonio, Artificial intelligence, large language models (llms), and retrieval-augmented generation (rag). new tools for accessing archival and bibliographic resources, Bibliothecae.It 13 (2024) 146–173. doi:10.6092/issn.2283-9364/19982.