Tracing the art market: a hybrid approach to digitise historical auction catalogues

Marilena Daquino[1], Francesca Mambelli[2], Valentina Pasqual[1], Valentina Rossetti[1,2,*] and Francesca Tomasi[1]

[1]Dept. of Classical Philology and Italian Studies, University of Bologna, via Zamboni 32, 40126 Bologna (Italy)
[2]Federico Zeri Foundation, University of Bologna, Convento di Santa Cristina, Piazzetta Giorgio Morandi 2, 40125 Bologna (Italy)

Abstract

Cultural institutions interested in the art market and the history of collecting increasingly face complex challenges in digitising, analysing, and disseminating auction catalogues. While digitisation initiatives exist, systematic evaluations of workflows for digitisation, transcription, segmentation, and review remain limited, making generalisation difficult. The production of fully interoperable Linked Open Data is rare, with semantic enrichment often limited to small data subsets. The Federico Zeri Foundation has launched a project to digitise 1,900 auction catalogues spanning 1869–1929 across six countries. This effort employs an iterative workflow combining automated rule-based and LLM methods with human validation to create semantically enriched representations. Acknowledging that digitisation choices influence historical interpretation, an incremental data-modelling strategy is adopted to accommodate variable transcription and segmentation quality. This article details such workflow and provides a preliminary evaluation of its effectiveness.

Keywords

Auction catalogues, digitisation workflow, ontologies

1. Introduction

An increasing number of cultural institutions are embracing the management of complex issues related to the digitisation, analysis, and dissemination of their sources. A particularly challenging endeavour is the digitisation of printed auction catalogues, which are primary sources of the history of the art market. Auction catalogues describe public sales of artworks and other collectables, documenting artworks' provenance, collectors' practices, market trends, and taste evolution [1]. While a few digitisation projects specific to auction catalogues show progressive methodological and technical sophistication (German Sale project [2], Frick Collection [3, 4], Mediate [5, 6], AuCaSe[1], Data Catalogue [7]), evaluations of proposed workflows for digitising, transcribing, segmenting, and reviewing data are missing or are still in their infancy, offering little help for being generalised. Moreover, projects rarely produce 5-star interoperable Linked Open Data, or semantic enrichment addresses only a small portion of it (Mediate [5, 6], Data Catalogue [7]), reducing data reusability and hampering the research questions (RQ) that data can answer.

In this article, we introduce and preliminarily evaluate a workflow for digitising, transcribing, and transforming data into Linked Open Data of Federico Zeri's auction catalogues collection. The corpus, including 1,900 ancient auction catalogues, is part of a larger collection of 36,000 catalogues and is representative of the international art market between 1869 and 1929, including volumes published in six different countries (Italy, France, Germany, the Netherlands, Great Britain and the USA). The digitisation project requires an iterative workflow that combines automated approaches, leveraging both

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

*Corresponding author.
marilena.daquino2@unibo.it (M. Daquino); francesca.mambelli6@unibo.it (F. Mambelli); valentina.pasqual2@unibo.it (V. Pasqual); valentina.rossetti7@unibo.it (V. Rossetti); francesca.tomasi2@unibo.it (F. Tomasi)
0000-0002-1113-7550 (M. Daquino); 0000-0002-2341-4375 (F. Mambelli); 0000-0001-5931-5187 (V. Pasqual); 0009-0007-5385-253X (V. Rossetti); 0000-0002-6631-8607 (F. Tomasi)

[1]https://aucase.inha.fr/, https://github.com/raphaelBarman/aucase-inha

rule-based systems and LLMs integrations, with human revision, resulting in a semantically enriched view of original data. Notably, our approach to data modelling recognises that digitisation choices shape results and, consequently, their historical interpretation. To this end, we propose an incremental approach to data modelling that accounts for different expected output scenarios based on the quality of transcription and segmentation.

2. Related Work

The digitisation of auction catalogues presents significant challenges beyond simple optical character recognition (OCR). Historical auction catalogues suffer from a variety of issues that compromise OCR performance, such as ink bleeding, image contrast variation, background noise and fungal degradation due to long storage times [8, 9]. Empirical studies show that OCR accuracy on similar materials ranges from 81% to 99%, with variability strongly influenced by the quality of the digitised image and the binarisation techniques applied [10, 11]. In addition, the layout of auction catalogues affects text segmentation [12]. Different languages are associated with different, unusual fonts, each presenting typographical variations that generate OCR errors [13].

Significant improvements in the processing of historical documents are offered by machine learning and LLMs, which support OCR and segmentation tasks. The AuCaSe project applies deep learning to image segmentation on 1,911 auction catalogues, using neural networks trained to recognise and separate blocks of text and images within pages. The Data Catalogue project [7] uses layout-based segmentation as a core component and open standards for data publication (XML/TEI). However, research is still in an early phase. Methodologies are neither documented nor evaluated, limiting their reproducibility. Recent benchmarks [14] indicate that digitisation workflows that experiment with LLMs may pave the way for new workflows for preserving and enhancing cultural heritage data. However, a one-size-fits-all solution has yet to be developed.

Moreover, knowledge extraction results are often not validated. While crowdsourcing campaigns may be envisioned to review, correct, or enrich the extracted data [7], the produced data are often published in non-interoperable formats and following custom schemas, leaving little room to evaluate them with respect to RQs and their effectiveness in scholarly enquiry. Efforts to produce structured data have focused on structure and layout (*e.g.*, METS, ALTO, SegmOnto[2]) [7, 15, 16, 17], whereas few projects have successfully addressed the use of ontologies to describe content (*e.g.*, Europeana Newspaper[3]) [18, 19]. However, such projects consistently report substantial quality issues derived from OCR accuracy [10, 11]. Since the quality of the knowledge extracted may vary significantly depending on digitisation choices, data modelling is (or should be) also affected. Most de facto standards, like CIDOC-CRM [20], assume knowledge is stable, *i.e.*, defined before the transformation process, leaving little room for different levels of granularity that reflect the actual data accuracy, which can vary throughout the digitisation process. For instance, in the Zeri auction catalogues project, we expect a number of catalogues to be fully digitised and semantically enriched (e.g. with lot descriptions, artists, type of objects, dates), while others will only have higher-level detail if the transcription and extraction is not accurate enough (e.g. broader types associated to the whole collection of lots, not to single objects). To the best of our knowledge, no prior project has investigated, in one place, the methodological implications of how digitisation practices and data-modelling choices interact, nor how expectations and outcomes in the digitisation process influence modelling decisions and the resulting data in relation to RQs.

3. Methodology

The goal of this work is to define a coherent end-to-end workflow for the digitisation and semantic representation of historical auction catalogues, adopting a modular approach, driven by incremental,

[2] https://segmonto.github.io/pj/vocabulary/
[3] http://www.europeana-newspapers.eu/

scalable RQ analysis. This methodology covers the entire data life cycle, from acquisition to publication, in line with FAIR principles [21]. Novel aspects include 1) the usage of LLMs to accomplish tasks like OCR and rule-based methods for segmentation (to identify lot descriptions) and 2) an incremental data modelling process, based on RQs and their feasibility, to extend results of the previous phase according to a scalable criterion.

According to recent research [14], a document-sensitive strategy suggests that auction catalogues (high scan quality, moderate typographic clarity and simple layouts) are well suited for LLM-enhanced OCR and improved formatting without semantic distortion. Results of [14] constitute the baseline for our experiment. Therefore, we adopt a state-of-the-art transcription tool, *i.e.* docling [22], as our first choice (v2.61.2, cpu-only version for the sake of the evaluation). To design the ontology, we adopt established methodologies grounded in Competency Questions [23] and ontology reuse practices [24]. Within the cultural heritage community, CIDOC-CRM and the Linked Art application profile [25] are widely regarded as gold standards. However, existing implementations offer limited support for representing and harmonising multiple levels of data granularity, which may not be available at different stages of a project. Our approach addresses this gap by envisioning four scenarios in which RQs can be answered, namely:

1. Using catalogue-level metadata (*e.g.*, Which catalogues are published in France?).
2. After digitisation and page classification (*e.g.*, Which catalogues contain illustrations?).
3. After full digitisation, transcription, segmentation, and knowledge extraction (*e.g.*, Which artists are mentioned in French catalogues?).
4. From catalogue-level metadata and extracted data with two levels of granularity (*e.g.*, What types of objects were sold at auction?).

This multi-layered perspective enables a flexible ontology that accommodates incremental data refinement without compromising semantic consistency. Our two parallel research lines dedicated to digitisation and transcription, and to knowledge graph generation, are summarised as follows.

3.1. Digitisation and transcription process

1. Analogue catalogues are digitised using professional scanners.
2. After digitisation, images are parsed by an LLM for page classification purposes (Pixtral 124B q16/4), to identify front matter, illustration, lot description, or blank page [26].
3. Digital images and LLM-generated classifications are stored on a dedicated IIIF server.
4. Images labelled as “lot description” are programmatically accessed via API to perform OCR, which returns a markdown file for each image, later collated into one for each catalogue.
5. Markdown files are parsed with rule-based segmentation mechanisms to retrieve 1) the list of lots, and 2) issues raised in the segmentation. Data are stored in CSV files.

3.2. Knowledge graph generation

6. We codesign RQs with art historians, to be used as CQs to design the ontology and validate data. We identified 27 RQs[4].
7. Questions are manually classified according to the above scenarios.

[4] https://github.com/dharc-org/zeri_auction_catalogues/blob/main/evaluation/Matrice_di_complessit%C3%A0_AuctionCatalogs-EN.xlsx

8. We design ontology patterns for the above CQs. Notably, the fourth scenario requires us to design two distinct patterns that complement each other. For instance, we design a pattern to associate the type of objects to the whole collection according to catalogue data, and a pattern to associate the type to single lots, as extracted by named entity recognition (NER)[5].

9. Catalogue metadata is transformed into RDF. Data address information such as authors, auction houses, places, dates, and seldom types and dates of objects[6].

10. We perform automatic data reconciliation to Wikidata to identify auction houses and authors. Recommended matches were manually revised and integrated into the RDF data.

11. Transcriptions of lot descriptions are revised by humans via a web application to correct segmentation, order, and text of lot descriptions.

12. Lot descriptions are converted into RDF following patterns defined in the fourth scenario.

13. NER is eventually performed to extract artists and object types from lot descriptions. Named entities are post-processed and reconciled to Getty ULAN and AAT [27]. Revised reconciliations are ingested in the graph.

From the RQ analysis, stems the decision to transcribe only pages labelled as "lot description", since most answers to scholars' questions are either included in catalogue metadata or in lot descriptions, while information included in paratextual sections (e.g. preface, conditions of sale) have lower priority. Moreover, RQs/CQs drive the ontology validation process. Activities 11-13 are planned, not yet implemented. The resulting graph populates the platform for data access. The semantic model will be the cornerstone for performing sophisticated queries on catalogues (*e.g.*, "which auctions in 1900 featured paintings attributed to Caravaggio?").

4. Results and preliminary evaluation

Results encompass the preliminary assessment of 1) LLM-assisted page classification, 2) rule-based segmentation methods, and 3) data modelling strategies. At this stage, we do not validate the quality of OCRed text, since we do not expect to include full-text lot descriptions in the final knowledge graph. Rather, in future works, we will evaluate the quality of the NER (still to be implemented and out of scope for this preliminary evaluation). Results are available online[7].

LLM-assisted page classification and rule-based segmentation methods. We validate our results on a sample of 20 catalogues (1%) written in three most representative languages (Italian, French, German) and with different, most common layouts (with and without illustrations and different lot description layouts). We manually assess the classification and segmentation results for pages classified as lot description and calculate precision and recall. We do not evaluate other page types' classification (*e.g.*, front matters, blank pages), nor perform segmentation on pages other than the ones classified as lot description. We consider a lot description correctly segmented when it includes the lot number and at least the title of the lot.

The page classification performs well in terms of recall ($r = 0.99$), while it struggles in precision ($p = 0.67$), which affects the precision of segmentation methods ($p = 0.94$). The low precision is due to a systematic misinterpretation in the page classification. Indeed, 526 of 573 pages classified as false positives (FP) include an illustration. The illustration usually includes a caption referencing the lot number of the artwork depicted and seldom a short description, which the classifier interprets as a lot

[5] https://github.com/dharc-org/zeri_auction_catalogues/blob/main/evaluation/PatternOntology_AuctionCatalogs.png
[6] https://github.com/dharc-org/zeri_auction_catalogues/blob/main/Zeri_cataloghi_RDF.ipynb
[7] https://github.com/dharc-org/zeri_auction_catalogues.git

Table 1
Precision and recall of classification and segmentation methods.

	Expected	Retrieved	FN	FP	Precision	Recall
Page classification	1194	1762	5	573	0.67	0.99
Page classification revised	1194	1236	5	47	**0.96**	**0.99**
Lot segmentation	13350	13798	449	781	0.94	0.97
Lot segmentation refined	13350	13537	449	520	**0.96**	**0.97**

Table 2
RQs, feasibility of knowledge extraction and ontology patterns.

Research line	#RQ	Method	Priority	Feasibility
Lot description	2	KE (2)	High (2)	High(1), Low(1)
Lot classification	8 (9)	KE(7), CD(1), both(1)	High (9)	High(2), Medium(4), Low(3)
Artist attribution	5	KE(4), CD(1)	High(3), Low(2)	Medium (2), Low(3)
Market trends	4 (6)	KE(1), CD(1), both(2)	High(3), Low(2)	High (2), Medium(3)
Historical insights	5	KE(5)	High(2), Low(3)	High(1), Medium(1), Low(3)

description, hence the misclassification. Such false-positive pages, when segmented to retrieve the list of lot descriptions, return a high number of FP (781), which include figure captions.

However, in our project, illustrations consistently appear in dedicated pages, and never together with the description of lots in the same page. Therefore, we can programmatically ignore illustration pages. The page classifier is refined to return a classification of "lot description" pages and a boolean value flagging page including an illustration. By pruning such pages, the number of false negatives lots (FN) does not change (449), while the FP sensibly decrease (781 to 520), and the precision of page classification significantly improves (before $p = 0.67$; after $p = 0.96$), as well as the precision of segmentation (before $p = 0.94$; after $p = 0.96$). Error analysis reveals that 449/13350 lots are omitted (3.4%), while 520/13350 lots (3.9%) include incorrectly splitted lots, duplicated lots, or lots with wrongly recognised lot numbers. We are aware that this naive result may not generalise to other auction catalogue collections, where the number of FN may increase.

Data modelling strategies. The RQs co-designed with cataloguers and art historians are summarised in Table 2. RQs have been first classified into 5 broad art historical research lines, namely: 1) lot description, 2) lot classification, 3) artist attribution, 4) auction houses, and 5) historical insights. Moreover, we categorised them according to the necessary or preferred method for retrieving the answer, *i.e.* via catalogue data (CD), knowledge extraction (KE), or both. Questions that could be answered with a different granularity level using both methods, *i.e.* using catalogue metadata only and/or with knowledge extraction, have been split into two subquestions, each requiring a different ontology pattern. The final number of RQs is shown in column "#RQ" whenever the column "Method" includes the value "both". Lastly, each question has been assigned a priority score (low to high) based on its expected impact on scholarly enquiry and a feasibility score that accounts for the expected success of knowledge extraction.

Notice that if an RQ could be answered using catalogue data (*e.g.*, Where did the auction take place?), we prioritised this method in classifying the RQ, since we expect higher accuracy, albeit lower recall. In these cases, knowledge extraction will complement incomplete information.

The result is a set of 27 RQs and corresponding ontology patterns. Out of the 27 RQs, in the first release of the knowledge graph, we plan to answer 15 questions, *i.e.*, 5 from catalogue data and 10 from knowledge extraction. Three of the 15 RQs can be answered using both catalogue data and knowledge extraction, namely: RQ1) Which object types are sold the most? RQ2) From what periods are the most sold works? RQ3) Which auction houses are associated with sales of specific object types/objects of

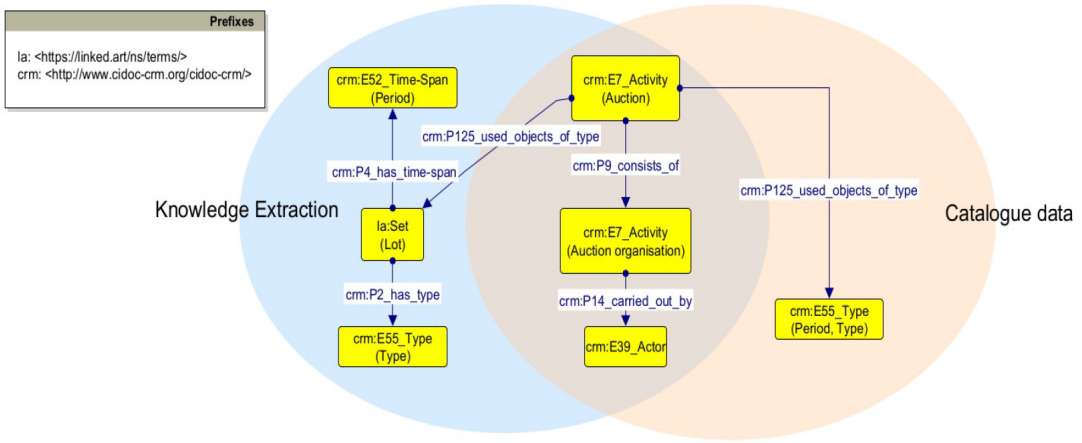

Figure 1: Venn diagram showing ontology patterns to represent data from knowledge-extraction (blue) and from the catalogue (orange). The centre shows their shared ontology patterns. In particular, RQ1 modelled with the core classes crm:E7_Activity (Auction), la:Set, and crm:E55_Type; RQ2 extends RQ1 by adding crm:E52_TimeSpan for temporal reference; RQ3 extends RQ1-2 with crm:E7_Activity (Auction organisation) and crm:E39_Actor for describing involved auction houses.

a certain period? In these cases, we designed two different ontology patterns according to the two granularity levels, as shown in Figure 1.

5. Discussion and conclusion

The preliminary results of this study highlight both the potential and the limitations of a hybrid workflow that combines automated methods, LLM, and human revision for the digitisation and semantic enrichment of historical auction catalogues. While our approach demonstrates that significant portions of the digitisation pipeline can be streamlined through automation, it also underscores that human expertise remains essential, particularly in semantic validation.

LLM-assisted page classification methods produce encouraging results when dealing with catalogue layouts that exhibit moderate visual regularity. This aligns with recent findings [14]. However, classification errors tend to cluster around visually ambiguous pages—especially those containing illustrations accompanied by captions referencing lot numbers. These errors have direct downstream consequences: misclassified pages are passed to the segmentation stage, thereby increasing the number of false positives. The segmentation of lot descriptions remains a critical bottleneck. Despite high recall in identifying lot boundaries, precision decreases in pages (1) with inconsistent typographic structures, such as indistinct numbering of lot descriptions, (2) erroneously classified as lot descriptions, and (3) where the LLM performing OCR hallucinates, *e.g.* duplicating or omitting lot descriptions, changing the markdown formatting of lists.

This finding aligns with longstanding difficulties in segmenting historical newspapers and similar materials [10, 12]. While rule-based approaches offer interpretability and reproducibility, they lack adaptability. Our long-term objective is to reduce the amount of human correction required after automated segmentation. However, the present results suggest that a fully automated segmentation pipeline remains unrealistic for heterogeneous historical collections. Instead, the most promising route may be to combine pattern-based heuristics with LLM-driven validation steps—using the model not to segment text directly, but to check the plausibility of rule-based segmentation (*e.g.*, by identifying if a text block contains lot's description essential features).

Although the current evaluation does not include NER, this step is expected to introduce further complexity. Our experience indicates that automatic matching performs unevenly across entity types and human oversight remains necessary to correct ambiguous or erroneous mappings. Future experiments will include the use of vector embedding-based similarity search, relying on open-weights models to

select reconciliation candidates from external authorities, and LLMs to re-rank and filter the candidates. The custom web interface is aimed at mediating, allowing fine-grained corrections that would be difficult to achieve with fully automated solutions, although it raises questions about scalability. Future work will evaluate the trade-off between accuracy and effort to reduce manual revision without compromising data reliability.

One of the main contributions of this work is the incremental approach to data modelling. Traditional cultural heritage standards assume relatively stable knowledge structures, which may be too rigid for corpora in which information emerges gradually and unevenly along the digitisation pipeline. Our four-scenario model reconciles data complexity with practical constraints, enabling the coexistence of catalogue-level metadata, partially processed catalogues, and fully enriched catalogues within the same semantic framework. This flexibility proved essential when classifying RQs and determining which could realistically be answered from early outputs (*e.g.*,metadata) and which required full transcription and extraction. The fact that several questions can be addressed at multiple granularity levels exemplifies the value of this approach: rather than adopting a monolithic data model that obscures uncertainty, we model uncertainty explicitly by providing alternative ontology patterns for different levels of completeness. This mirrors emerging practices in digital humanities, where uncertainty is increasingly recognised as a meaningful dimension of knowledge representation [28, 29].

The results of the RQs analysis demonstrate that many high-priority questions can be answered even with incomplete transcription, supporting the prioritisation of lot-description pages and the deferral or reduction of transcription for paratextual sections. This confirms that aligning digitisation decisions with scholarly needs is not only desirable but necessary to ensure efficient resource allocation.

In summary, the findings highlight several broader implications that extend beyond the Zeri collection. First, digitisation workflows should be designed as adaptive pipelines, capable of adjusting to the heterogeneity of historical materials rather than enforcing rigid uniformity. Second, the integration of LLMs into heritage workflows should be guided by careful evaluation: while LLMs offer significant advantages, they are not universally reliable and must be embedded within controlled processes. Third, this project illustrates the value of aligning digitisation strategies with RQs from the outset, ensuring that data is meaningful for scholarly inquiry. Finally, the approach presented here contributes to a much-needed discussion on how digitisation choices shape the historical research questions. By explicitly modelling different levels of granularity and by documenting how data is generated and corrected, we offer a transparent framework that others can reuse or adapt. This study represents a preliminary evaluation based on a small subset of the corpus. Future work will expand the evaluation to a larger sample and include the assessment of OCR accuracy, NER quality, reconciliation precision, and end-to-end reproducibility.

Declaration on Generative AI

Authors used GPT-4 for Grammar and spelling check. Afterwards, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

Acknowledgments

This research is partially funded by HORIZON Europe Research and Innovation Action, project LUMEN "Linked User-driven Multidisciplinary Exploration Network" GA ID: 101187940.

References

[1] P. Fletcher, A. Helmreich, Digital Methods and the Study of the Art Market, in: The Routledge Companion to Digital Humanities and Art History, 2020.

[2] C. Huemer, The "German Sales 1930–1945" Database Project, Collections: A Journal for Museum and Archives Professionals 10 (2014) 273–278. doi:10.1177/155019061401000306.

[3] K. West, C. Llewellyn, J. Burns, Automatic Processing and Segmentation of Auction Catalogs, in: Digital Humanities 2010, 2010, p. 267.
[4] G. Nadasky, Preserving web-based auction catalogs at the Frick Art Reference Library, D-Lib Magazine 20 (2014). doi:10.1045/march2014-nadasky.
[5] A. C. Montoya, Middlebrow, Religion, and the European Enlightenment: A New Bibliometric Project, MEDIATE (1665–1820), . URL: https://hdl.handle.net/2066/183283.
[6] A. C. Montoya, M. Hulsbosch, H. Blom, E. Chayes, A. De Wilde, R. Jagersma, J. Reboul, J. Rozendaal, MEDIATE database, 2022. URL: https://mediate-database.cls.ru.nl/, ongoing.
[7] H. Scheithauer, S. Bénière, L. Romary, Automatic retro-structuration of auction sales catalogs layout and content, in: DH2024 - Reinvention and Responsibility, Alliance of Digital Humanities Organizations, Washington DC, 2024. URL: https://hal.science/hal-04547239.
[8] S. Bhowmik, K. Omar, M. F. Nasrudin, Degraded Historical Document Binarization: A Review on Issues, Challenges, Techniques, and Future Directions, Applied Sciences 11 (2021) 6984. doi:10.3390/app11156984.
[9] K. Saddami, et al., Effective and fast binarization method for combined degradation on ancient-documents, Journal of Electrical Systems 20 (2022) 3556–3568. doi:10.1016/j.heliyon.2019.e02613.
[10] B. Gatos, D. Danatsas, I. Pratikakis, S. J. Perantonis, Europeana Newspapers OCR Workflow Evaluation, in: Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing (HIP '12), 2012, pp. 90–97. doi:10.1145/2809544.2809554.
[11] R. Holley, How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs, D-Lib Magazine 15 (2009). URL: https://www.dlib.org/dlib/march09/holley/03holley.html.
[12] J.-Y. Ramel, et al., User-driven page layout analysis of historical printed books, International Journal on Document Analysis and Recognition 9 (2007) 243–261. doi:10.1007/s10032-006-0018-9.
[13] Eureka Network, Digitising historical documents, 2024. URL: https://www.eurekanetwork.org/impact/digitising-historical-documents/.
[14] O. M. Machidon, A. L. Machidon, Comparing OCR Pipelines for Folkloristic Text Digitization, arXiv preprint arXiv:2507.19092, 2025. doi:10.48550/arXiv.2507.19092.
[15] Library of Congress, METS: Metadata Encoding and Transmission Standard, 2024. URL: https://www.loc.gov/standards/mets/.
[16] Library of Congress, ALTO: Technical Metadata for Optical Character Recognition, 2024. URL: https://www.loc.gov/standards/alto/.
[17] Bibliothèque nationale de Luxembourg, METS/ALTO Technical Guidelines for Newspaper Digitisation, 2023. URL: https://data.bnl.lu/data/historical-newspapers/mets-alto/.
[18] C. Neudecker, et al., Europeana Newspapers: A Gateway to Historical Digitised Content, EuropeanaTech Insight, 13, 2019. URL: https://pro.europeana.eu/page/issue-13-ocr.
[19] G. M. Silva, A. C. Terra, Cultural heritage on the Semantic Web: The Europeana Data Model, Information 14 (2023) 109. doi:10.3390/info14020109.
[20] CIDOC Documentation Standards Working Group, CIDOC Conceptual Reference Model (ISO21127:2023), 2023. URL: https://cidoc.mini.icom.museum/.
[21] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, et al., The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018. doi:10.1038/sdata.2016.18.
[22] N. Livathinos, C. Auer, et al., Docling: An efficient open-source toolkit for ai-driven document conversion, arXiv preprint arXiv:2501.17887, 2025. doi:10.48550/arXiv.2501.17887.
[23] V. Presutti, E. Daga, A. Gangemi, E. Blomqvist, extreme design with content ontology design patterns, in: Proc. Workshop on Ontology Patterns, 2009, pp. 83–97. URL: https://dl.acm.org/doi/10.5555/2889761.2889768.
[24] V. A. Carriero, M. Daquino, A. Gangemi, A. G. Nuzzolese, S. Peroni, V. Presutti, F. Tomasi, The landscape of ontology reuse approaches, in: Applications and practices in ontology design, extraction, and reasoning, IOS Press, 2020, pp. 21–38. doi:10.3233/SSW200033.

[25] Linked Art Community, Linked Art: A Web Standard for Describing Art, and the Right Way to Describe It, 2024. URL: https://linked.art/.
[26] P. Bonora, T. Vitale, AI in-the-loop: the AI-assisted digitisation workflow of the Fondazione Zeri's auction catalogues collection, 2026. In Proceedings of IRCDL 2026 [forthcoming].
[27] Getty Research Institute, Art & Architecture Thesaurus Online, 2024. URL: http://www.getty.edu/research/tools/vocabularies/aat/.
[28] M. Daquino, F. Tomasi, Historical Context Ontology (HiCO): A Conceptual Model for Describing Context Information of Cultural Heritage Objects, in: Metadata and Semantics Research: 9th International Conference, MTSR 2015, Proceedings, Springer, 2015, pp. 424–436. doi:10.1007/978-3-319-24129-6_37.
[29] J. Flanders, F. Jannidis (Eds.), The shape of data in digital humanities: Modeling texts and text-based resources, Routledge, 2019.