✍️ Report

Course:	Embedded Research Project – Masters on Display: Navigating 19th-Century Dutch Art Exhibitions
Instructor(s):	Dr. H. (Houda) Lamqaddam
Assignment:	ERP Report
Name:	Ava Zandieh Doulabi (13439790) Maria Elpiniki Zafeiraki (15446891) Pascale Stomp (13492667) Yağmur Sarıgül (15698580)
Date:	25.04.25

Navigating 19th Century Dutch Art Exhibitions:

An OCR and LLM Approach to Text Extraction and Structuring in Digital Art History

Abstract

This report addresses a major challenge faced by researchers in the digital humanities working with large volumes of historical textual sources that remain underexplored due to the intensive manual labor required to process them. Although computational methods have been utilized in digital art history and its subfields to analyse text or images, the extraction and structure of textual data from archival sources with state-of-the-art AI tools has not yet been extensively studied. Nowadays, with the continuous improvement of OCR technologies and LLMs, new opportunities to automate these enormous manual processes have emerged to make archival data more accessible and researchable. For this project, we explored OCR Engines and LLMs to extract and structure data from the catalogs of 19th century Dutch art exhibitions called Living Masters at the RKD. In this paper, we present the preprocessing and structuring steps of the tools we experimented with and reflect on their potential to support research within the field of digital art history.

1. Introduction

In an era defined by rapid digitization and the sharp, continuous growth of computational tools, cultural heritage institutions—Galleries, Libraries, Archives, and Museums (GLAM)— are increasingly confronted with a paradox; despite the extensive efforts to digitize large bodies of historical archives, manuscripts, and exhibition catalogues, much of this rich historical data remains locked away in unstructured formats, inaccessible for deeper research and public engagement (Sotirova et al. 2012).

Instead of a merely archival practice, digitization has become, nowadays, an essential step transforming how we understand, analyze, and interact with historical documents (Poulopoulos and Wallace 2022). As GLAM institutions continue collecting and digitizing extensive records, there is an emerging need to shift through and extract meaningful information to support digital curation and interoperable formats such as Linked Open Data (Chiquet et al. 2024). In this regard, automated data extraction using Optical Character Recognition (OCR) and Large Language Models (LLMs) offers a promising solution to bridge this gap efficiently while maintaining accuracy and achieving accessibility by converting vast, unstructured digital archives into well-organized, searchable data formats (Fleischhacker et al. 2024; Hyvönen 2020).

Automated data extraction can revolutionize how cultural heritage collections are used. Manual transcription of historical documents is not only labor-intensive and slow but also prone to errors, especially when deciphering archaic typefaces or handwritten texts (Hoetjes 2025). In contrast, modern OCR systems, especially when enhanced with machine learning, can process large volumes of historical material far more efficiently (Malladhi 2023). Such advancements significantly reduce processing times, allowing experts to engage in deeper analysis and interpretation.

Beyond efficiency, accuracy is a critical advantage of state-of-the-art automated extraction methods. New generations of computational tools—like those emerging from recent research into historical document processing—have shown remarkable outcomes. For example, Khan et al. (2024) highlight that state-of-the-art OCR approaches enhanced with deep learning can accurately transcribe historical texts that were once nearly indecipherable. Moreover, the recent CalliReader project (Luo et al. 2025) further underscores how embedding-aligned vision-language models can contextualize complex scripts such as Chinese calligraphy with accuracy that measures up to human experts.

Improved data extraction not only increases efficiency and accuracy but also significantly enhances accessibility. When unstructured digitized content is converted into structured, searchable databases, cultural heritage becomes a more powerful resource. Enhanced accessibility benefits a wide range of stakeholders—from scholars and researchers to the general public—enabling interdisciplinary research and fostering cultural preservation (Lian and Xie 2024). This transformation supports the development of digital archives that serve as both repositories of knowledge and dynamic platforms for scholarly inquiry.

The RKD (Netherlands Institute for Art History) houses one of the largest collections of art historical materials in the world (2025), including the comprehensive series of exhibition catalogues from the Exhibitions of Living Masters (Tentoonstelling van Levende Meesters), which documented contemporary art in the Netherlands for more than a century. As part of the Metamorfoze program for the preservation of Dutch cultural heritage, 877 catalogues have been digitized, listing between 100-600 artworks each (Koot and Kapelle 2015). Despite digitization, their vast volume makes the catalogues' contents inaccessible for in-depth analysis and makes it difficult to connect them with the RKD's artists and artworks databases. Implementing an OCR and LLMs-driven workflow could unlock the rich contextual data in these texts, improving their exploration prospects and serving as a prototype for similar projects or institutions.

However, integrating these technologies does not come without challenges. Automated extraction from historical texts often involves handling numerous technical complexities, such as variable image quality, deteriorated texts, or inconsistent layouts, and that is why it is important for targeted preprocessing and prompting practices to be designed (Fleischhacker et al. 2024). Additionally, performance evaluations of OCR tools and commercial LLMs on historical documents reveal considerable variability in accuracy and reliability (Khan et al. 2024). These challenges lead to critical research questions, which we will address in the following sections:

- **How can OCR and LLMs be optimally integrated to support automatic data extraction from historical texts?**
- **How can non-standard data be effectively extracted and structured?**

The practical implications of these research questions can offer valuable insights regarding faster and more reliable data processing of archival material and enhance the research value of cultural heritage collections, creating a substantial impact on digital humanities, as well as archival and museum studies. Additionally, this study aims to contribute to existing literature on digital art history, computational text extraction and structuring from historical documents and specifically, art exhibition catalogues, and implementation of AI tools by offering a methodological pipeline and comparative performance analysis of OCR Engines and

LLMs, which is tailored to the Living Masters case study but could serve as a reference for similar initiatives in GLAM institutions.

This paper is organized into several key sections. In the first section, we review relevant literature on digital art history viewpoints and practices, text extraction for digital humanities, and specifically, for catalog archival material and respective OCR and LLMs applications in similar projects. After providing the necessary contextual framework on the Exhibition of Living Masters catalogs, highlighting their significance, and describing their format and categorization, the methodology section focuses on the extraction, structuring, and output cleaning steps that were followed, as well as a comparative summary of OCR and LLM tools that we experimented with to achieve the project's objective. In the results part, we reflect on the process, present our findings, and note certain challenges, whereas in the discussion and limitations, we emphasize both the transformative potential and the ethical and practical implications of such workflow. Finally, we suggest points and practices for future research, scalability, and the project's next steps. The pipelines' Python code scripts, along with examples and instructions for implementation, can be found in a public GitHub repository.

2. Literature Review

2.1. Digital Art History

The emergence of digital art history (DAH) as a scholarly field has raised several inquiries about how digital tools and methodologies reshape the epistemological foundations, interpretive practices, and collaborative structures of art historical research. As new thinking avenues and fields of activity have arisen from artwork digitization to the prioritization of "increasingly rich, user-friendly databases, and online publications" (Drucker et al. 2015), the significance of taxonomies and data, and format standardization has become evident to allow for extensive sharing of digital files across platforms or users and facilitate macro as well as micro-analysis methods (Drucker et al. 2015). Johanna Drucker's 2013 distinction between *digitized* and *digital art history* has served as a key reference point for framing the field, but also as a compass that has guided the ongoing scholarly discourse. However, in a more recent paper by Rodríguez-Ortega (2019), DAH is perceived as a methodological and epistemological extension of "traditional" art history, and it is emphasized that the term itself and its practices should be approached critically in light of the post-digital society and post-humanism. Impett and Offert (2022) revisit Johanna Drucker's foundational distinction between digitized and truly DAH, arguing that a genuinely "digital" art history has begun to take shape with the advent of multimodal machine learning models such as CLIP. These models, capable of interpreting both text and image data, enable new forms of analysis that move beyond object recognition and categorization. They call for DAH to embrace cross-disciplinary dialogue with media studies, critical AI studies, and computational humanities, highlighting that the field's future lies not just in adopting new tools, but in critically engaging with the epistemological frameworks those tools entail.

It has also been previously cautioned that such digital practices must be critically examined in a reviewed epistemological and methodological art history framework (Rodríguez-Ortega 2013), as digitization is not a neutral representation but a form of interpretation, and computational tools should not be presumed objective (Drucker 2013). Moreover, Bishop (2018) in her piece with the indicative title "Against Digital Art History" argues that DAH's reliance on quantitative methods deprives the empirical and philosophical depth of analysis in art and turns abstract, qualitative research questions into data and statistics as an outgrowth of neoliberalism.

On the other hand, Manovich (2015) introduces foundational data science concepts—such as features, feature space, and dimensionality reduction—to advocate for scalable, quantitative analysis of cultural artifacts, whereas Baca, Helmreich, and Gill (2019) support that digital methods are no longer peripheral but central to art historical research, teaching, and publishing. They highlight three key areas—databases, computational analysis, and digital publishing—and emphasize that digital tools require critical, historically informed engagement, but also collaborative work, transparency, and a reconsideration of academic norms to make the field more inclusive and responsive to contemporary challenges. Fisher and Swartz (2014) reflect on THATCamp CAA 2014 as a pivotal moment in DAH, emphasizing how digital practices are becoming integral to research, teaching, and collaboration in the field. They highlight the importance of open knowledge sharing, institutional recognition of digital work, and the need to equip scholars—especially emerging ones—with digital literacy and coding skills. Rather than positioning the digital as a solution to disciplinary challenges, they argue it should be understood as a tool for reimagining and expanding traditional methodologies. Similarly, Jaskot (2019) emphasizes the importance of situating digital methods within the social history of art, arguing for interdisciplinary approaches that address art's socio-political contexts across scales. Complementing these perspectives, Zweig (2015) offers a genealogical critique of DAH's perceived novelty, revealing how art historians have long engaged with digital thinking and methodologies. Finally, Brey (2021) focuses on resources, tools, and training for DAH students and practitioners alongside publications that showcase the potential of advanced digital methods of computer vision in image analysis.

These contributions underscore that DAH does not refer to a unified methodological practice, but an evolving field at the intersection of technological innovation, critical theory, and disciplinary self-reflection. Similar debates have emerged in computational literary studies (CLS), a DAH subfield that likewise engages with the promises and challenges of digital methods in the humanities. However, CLS practitioners have also developed a more mature and specialized toolkit for textual analysis, serving as a methodological reference point for other fields seeking to work with large-scale or complex textual data. For example, CSL utilizes techniques such as text mining (Dilai and Dilai 2024), topic modeling (Storm and Rainey 2024; Chu, Keikhosrokiani, and Asl 2022), vector semantics (Sobchuk and Šeļa 2024), stylometry (Scott Alen 2021; Salgaro 2023), and AI and machine learning (Slocombe and Liveley 2024; Ros, Bjarnason, and Runeson 2017). At the same time, scholars in CLS have explored the complementarity of distant and close reading (Amangazykyzy et al. 2025; Cuppen and van den Ven 2019) and the importance of critical but collaborative engagement with algorithms, corpora selection, and metadata structure (Bode and Bradley 2024; Yadav 2024).

DAH and related subfields such as CSL have largely concentrated on analytical frameworks or specific methodologies for text analysis, however, it remains rare to find projects that focus on pre-analysis workflows—that is, on making such historical documents more machine-readable and accessible before the interpretation part. This project aims to contribute to DAH by providing such a methodological example, focusing on the extraction and structuring of text from digitized art exhibition catalogues.

2.2. Text Extraction from Art Exhibition Catalogues and Historical Texts Using OCR

Art exhibition catalogues—whether originally digital or later digitized—constitute a distinct and valuable corpus for art historical research (Brey 2021). As early as 1975, the issue of seamless access to them through consistent cataloguing using computational methods has been raised (Smith and Treese 1975). Meanwhile, more recent research has focused on the potential for (near) fully automated conversion of PDF documents of digitized catalogues into semantically enriched, structured data. Gabay et al. (2021) present ways to enhance the Artl@s project database, addressing challenges such as text recognition, data standardization, and the creation of searchable databases. Additionally, Scheithauer, Bénière, and Romary (2024) discuss the automatic retro-structuration of auction sales catalogs, utilizing layout segmentation and information extraction technologies such as Name Entity Recognition (NER). Their method combines layout segmentation—automatically detecting catalog sections like front matter or entry lists—with fine-grained annotation using named entity recognition (NER) and text labeling models.

Although specific references to art exhibition catalogues are relatively scarce in the existing literature, comparable studies have been conducted regarding bibliographical records or dictionaries that have a similar structure. Goes (2019) presents a comprehensive study on the application of text mining techniques to structure OCR output from historical Brinkman catalogues. The study addresses a key challenge in digital humanities: transforming large-scale, print materials into machine-readable formats, by proposing a multimethod pipeline beginning with manual and semi-automated preprocessing to correct layout recognition errors and segment entries, followed by the extraction of bibliographic metadata such as author, title, publisher, and year. Likewise, Maxwell and Bills (2017) describe the ongoing efforts to use OCRed text for the creation of structured lexicons, making note of common errors and the need for standardized formatting. Through their approach, they emphasize the importance of preserving linguistic data by making digital formats of printed resources accessible and usable for research and language preservation. Ultimately, Bago and Ljubešić (2015) apply machine learning techniques to annotate language and structure in an 18th century dictionary by utilizing systematic patterns to train models and addressing challenges associated with digitizing historical lexicographic resources.

This reliance on annotated data highlights the broader challenges and methodologies within historical document processing (HDP), a field that encompasses a full pipeline from image preprocessing to semantic transcription and layout analysis. In their survey of the techniques, tools, and trends in HDP, Philips and Tabrizi (2020) highlight the role of annotated datasets and training data throughout the digitisation workflow, particularly in the context of preprocessing (e.g., binarisation, layout analysis), OCR, and handwritten text recognition (HTR). In their paper, they argue that while numerous algorithms exist, the scarcity of high-quality, annotated datasets remains one of the field's major bottlenecks, underscoring that advancements in HDP are tied to the availability of robust, domain-specific datasets that represent the historical and linguistic diversity of archival materials.

Building on these challenges, there have been notable efforts in the digital humanities that explored adaptable, user-friendly workflows that mitigate reliance on extensive annotated datasets while still addressing the complexities of historical document structure and OCR noise.

For example, Khemakhem et al. (2020) propose an information extraction workflow for digitised, entry-based documents like dictionaries and catalogues, using machine learning tools GROBID-Dictionaries and GROBID-Cat. The project addresses common challenges in digital humanities, such as OCR noise and data modeling, while aiming to make extraction tools accessible to non-technical users. Similarly, Weber (2021) presents a multimodal OCR pipeline tailored for digitising historical archival documents, addressing challenges such as low image quality, inconsistent layouts, and degraded text. The pipeline integrates document preprocessing (e.g., rotation, scaling, binarization) with OCR and post-OCR enhancement methods, including spelling error correction and temporal entity recognition. Finally, Mechi et al. (2020) present a method to extract full-text lines from archival documents by combining a U-Net-based model with a post-processing step. Tested on Tunisian archive data, the method achieved over 97% accuracy.

While these studies highlight major advances in OCR pipelines and structured extraction from historical documents, few have focused specifically on the hybrid use of OCR and AI models—especially LLMs—to extract text and structure art exhibition catalogues' processing metadata, which motivates the present study.

2.3. Beyond Extraction: Structuring Historical Texts' Extracted Data Using LLMs

The growing integration of AI tools and LLMs seems to have opened new possibilities for automating and enriching the analysis of historical documents. As a result, recent research across various disciplines has focused on integrating LLMs on top of OCR systems to enable more efficient text extraction and structural analysis in (historical) documents, particularly where degraded print, diacritics, and archaic language pose barriers to accurate transcription. Thomas et al., for example, (2024) demonstrate that instruction-tuned LLMs like Llama 2 can significantly reduce OCR errors by more than 50% in 19th century newspapers compared to earlier models like BART. Similarly, Do et al. (2025) use LLMs alongside reference ebooks to restore Vietnamese texts with missing diacritics and corrupted characters, outperforming existing spell-correction tools, whereas Veninga (2024) evaluates a fine-tuned version of the ByT5 model for post-OCR correction and finds that factors like text preprocessing and the size of input windows (e.g., 50 characters) can strongly affect the model's ability to fix errors.

In a more applied setting, Weber (2021) develops a semi-automated OCR pipeline that includes LLM-based spell correction and date extraction, designed to meet the needs of historians working with archival data. Abdellaif et al. (2024) go further by embedding LLMs into Robotic Process Automation (RPA) systems, showing that their system, LMRPA, completes OCR tasks up to 50% faster than existing platforms like UiPath. Lastly, Vangeli (2024) demonstrates how LLMs can act as powerful preprocessing tools, automatically cleaning and structuring unlabelled OCR text into datasets ready for machine learning, helping reduce the need for manual data cleaning in research workflows.

These studies highlight the growing potential of LLMs in enhancing OCR and automating document processing, focusing on well-structured texts or high-resource languages. In contrast, more complex, multilingual, and visually diverse sources—such as historical art exhibition catalogues—remain comparatively underexplored.

2.4. Bridging the Gap: Toward Structured Data from OCR and LLMs in the *Living Masters* Exhibition Catalogues

While digitization efforts of archival materials made them more available, many of them remain unstructured and inoperable for computational research or seamless integration to online databases. Even though fields such as DAH, CLS, or HDP have developed a variety of advanced tools for analyzing large-scale cultural datasets, they mainly rely on the availability of clean, structured text. As a few projects are starting to address the practical challenge of transforming noisy OCR output into machine-readable formats using AI tools, this project aims to fill a gap in the existing literature by proposing, applying, and evaluating a tailored OCR and LLM pipeline applied to the Living Masters exhibition catalogues.

Our goal is to make these texts more accessible, searchable, and interoperable by providing a practical workflow for GLAM institutions, which contributes methodologically to the growing field of DAH and art exhibition catalogue processing in the age of advanced LLMs. By emphasizing preprocessing, prompt engineering, and post-processing practices, we aspire to bridge the divide between technical innovation and critical engagement in digital humanities, offering a case study on how automated extraction and structuring can serve as a valuable tool for future research in digital art history, archival studies, and digital scholarship.

3. Living Masters Context

3.1. The Exhibitions of the *Living Masters*

The Exhibitions of Living Masters were a cornerstone of the 19th century Dutch art scene, organized in different cities of the Netherlands from 1807 to 1917 (Koot and Kapelle 2015). These exhibitions were first held in Amsterdam, Rotterdam, and The Hague and with time expanded to smaller cities like Haarlem, Dordrecht, Den Bosch, Utrecht, Arnhem, Leeuwarden, Groningen and more (Koot and Kapelle 2015). From emerging artists to already established ones, many Dutch artists in the 19th century showcased and sold their artworks from different disciplines without a commission fee (Koot and Kapelle 2015). These exhibitions provided a platform for Dutch artists to see the current artistic production from other cities and countries (Koot and Kapelle 2015). By bringing art collectors, artists, critics and art dealers together on one platform, the exhibitions contributed to the growth of art trade (Koot and Kapelle 2015). The Exhibitions of Living Masters also provided a significant platform for women artists. Despite their limited number and upper-class background, women were able to exhibit their work from the very first to the final exhibition (Koot and Kapelle 2015).

Compared to other countries like France (1665) and England (1770), the open public art exhibitions bloomed later in the Netherlands (Koot & Kapelle 2015). This delay in the exhibition scene in the Netherlands was due to the lack of a structured art academy and institutional governing body. With the foundation of Koninklijk Instituut van Wetenschappen and Letterkunde en Schoone Kunsten and with the initiation of Napoleon's brother Louis Bonaparte, the first Dutch king, the Exhibitions of Living Masters came to life (Koot & Kapelle 2015).

Following the Paris salon exhibition model, the newly established institutions organized the exhibitions backed by strong support from local communities and associations (Koot & Kapelle 2015). This created a uniquely decentralized organisation structure for the Dutch exhibitions that remained as such after the establishment of art academies and as the country's French rule came to an end (Sillevis 1991). What began as a single annual event quickly gained momentum and increased in frequency, popularity, and public demand. Crucially, the Living Masters art exhibitions filled a national artistic gap by being the first of their kind held in the Netherlands for a broader audience. Consequently, the exhibition series became the beginning of a new public art market with a new kind of art audience (Tibbe 2021).

3.2. Exhibition Catalogs

Exhibition catalogs are globally a staple of the art world for their function as knowledge disseminators as—both in contemporary and historical practice—they give insights into participating artists, artworks and are ideally accessible snapshots into often temporary art exhibitions (Joyeux-Prunel and Marcel 2015). They generally "vary from simple listings of works to extensive research tools" (Smith and Treese 1975, 471) and have historically become more diverse over the years. Some catalogs only list artworks, while others include illustrations, prices, additional venue information, or more in-depth details about artworks and artist portfolios or biographies. As such, calls have been made for more extensive methods to research catalogs in large volumes (Joyeux-Prunel and Marcel 2015).

In context of the Exhibitions of Living Masters, the catalogs remain an essential primary source that allow historians to reconstruct the evolution of the Dutch art scene in the 19th century (Koot & Kopelle 2015). However, since the catalogs were produced using low cost materials on poor-quality paper for the purpose of functionality, most of the surviving catalogs are currently in a deteriorated state (Koot & Kopelle 2015). Given the fragile condition of many catalogs, the RKD initiated conservation efforts in 2012 under the name of Bureau Metamorphoze. The project team collaborated with the Hoogduin restoration firm and Picturae to digitize all volumes and archive them in the RKD library as searchable PDF files (Koot & Kopelle 2015).

Over the years, the RKD has assembled the most extensive collection of the Living Masters exhibition catalogs currently holding approximately 877 digitized copies, 234 of which are unique editions, while the rest are made up of reprints, (handwritten) supplements, or printed addenda. The digitization has not only preserved the catalogs but also opened new avenues for research and revealed the need for deeper, structured access to their content.

Typically, the catalogs include the artist's name and city—sometimes even a complete address—along with the artwork's title, occasionally accompanied by a short description, poem, or notes on medium and genre. Additional features may include abbreviations denoting participation in competitions, previous awards or membership status to the hosting art association, a cross signals the recent passing of an exhibiting artist, while an asterisk generally indicates that a work was not for sale. Several catalogs also contain front and back-page price lists with corresponding currency values.

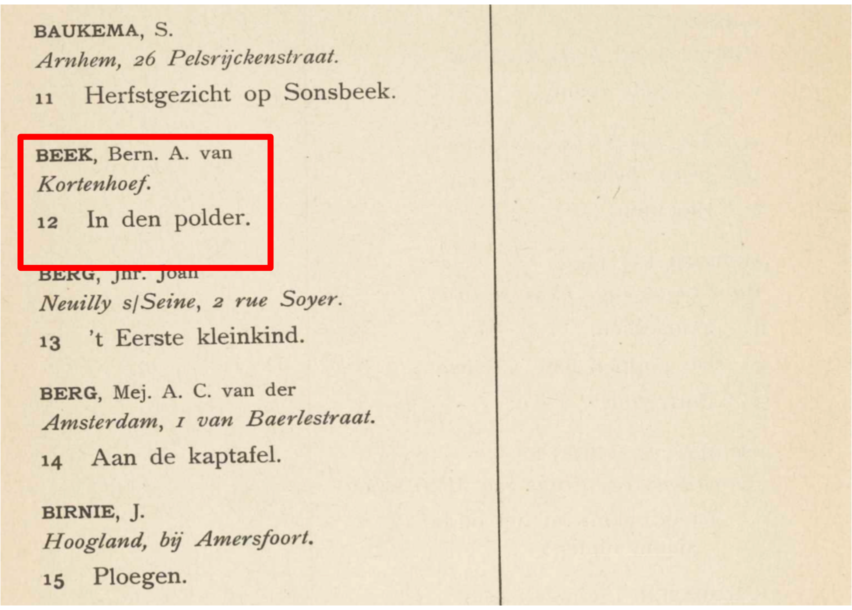
3.3. Standardization of the catalogs

Before extracting the data from the catalogs, we first had to decide what kind of information would be necessary. This was done in close collaboration with the RKD experts team and by first observing their existing Excerpts Database (RKD Research 2025), which hosts some manually handled catalogs. In this process, two standardization challenges were identified.

First, the manually extracted catalogs were not standardized to include the same categories—or even the same kind of data—in their respective catalogs, leading to inconsistencies in the existing database. To avoid this inconsistency issue in the automation process, the formats and structures of the catalogues were observed closely. As a result, it was noticed that the format and structure of the catalogs vary depending on the city, year, and organizing association (Figures 1 & 2). The variety in catalog formats has created some complications in the automation process. The key differences between formats that present challenges were as follows:

- Irregular Entry Numbering: Catalogue entries sometimes included letter-based numbers (e.g., 14a, 446b), making it difficult for models to maintain a consistent order.
- Inconsistent Currency Formats: Currencies were written in different forms in each catalog, including: f, gulden, mark, fr, and kr.
- Repetitive Entries and Shorthand Terms: Common shorthand such as "idem," "dito," or quotation marks (" ") were inconsistently applied to indicate repetition. This variation in shorthands creates difficulty for LLMs to accurately structure them with one customized prompt.
- Layout Variations: Layout inconsistencies included entries with and without street addresses, extra lines for ownership or material types, and different punctuation or formatting styles. Additionally, one of the major layout differences was the location of the artist's name. To illustrate, in some catalogs, the artwork name appears next to the artist's name, while in others, the artwork title is listed under the artist's name.
- Non-standard Formatting for Additional Info: Details such as medium/material (e.g., pastel, aquarel), location (e.g., Drenthe), or ownership were sometimes listed in parentheses and sometimes quoted, without consistency.

To account for these differences and enable structured data extraction, we aimed to develop a classification system of categories tailored to one general catalog type encountered throughout research.



BAUKEMA, S.
Arnhem, 26 Pelsrijckenstraat.
11 Herfstgezicht op Sonsbeek.

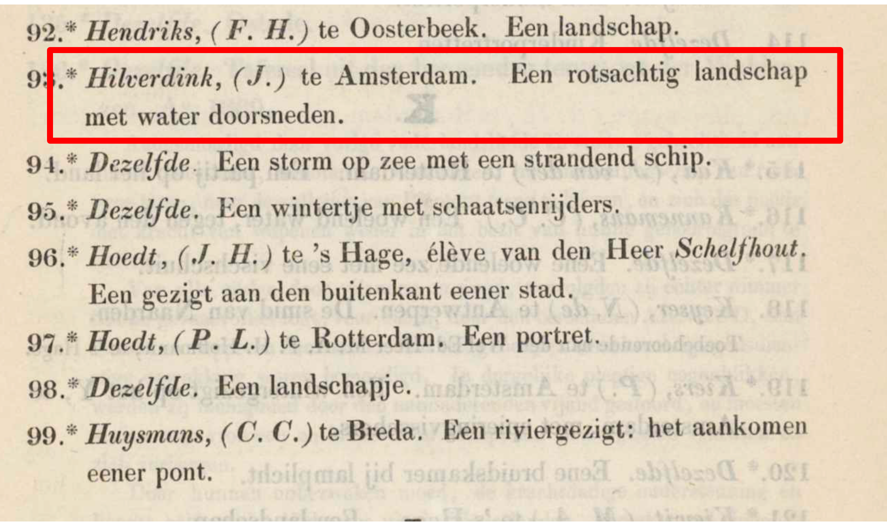
BEEK, Bern. A. van
Kortenhoef.
12 In den polder.

BERG, Jhr. Joan
Neuilly s/Seine, 2 rue Soyer.
13 't Eerste kleinkind.

BERG, Mej. A. C. van der
Amsterdam, 1 van Baerlestraat.
14 Aan de kaptafel.

BIRNIE, J.
Hoogland, bij Amersfoort.
15 Ploegen.

Figure 1.



92.* *Hendriks, (F. H.)* te Oosterbeek. Een landschap.

93.* *Hilverdink, (J.)* te Amsterdam. Een rotsachtig landschap met water doorsneden.

94.* *Dezelfde.* Een storm op zee met een strandend schip.

95.* *Dezelfde.* Een wintertje met schaatsenrijders.

96.* *Hoedt, (J. H.)* te 's Hage, élève van den Heer *Schelfhout.* Een gezigt aan den buitenkant eener stad.

97.* *Hoedt, (P. L.)* te Rotterdam. Een portret.

98.* *Dezelfde.* Een landschapje.

99.* *Huysmans, (C. C.)* te Breda. Een riviergezigt: het aankomen eener pont.

Figure 2.

Second, due to the linked nature of the overarching database, the specificity with which the catalogs were structured could not always be reflected in the most fine-tuned manner. Therefore, we aimed to design the standardized categorization of the data, using the linked nature of the database to its advantage.

3.4. Final categorization of the data

Fine-tuning categorizations by using the linked nature of the database to its advantage meant that some imagined categorizations such as artist gender, starting and end dates, exhibition titles or hosting cities could be ignored as these were already stored and could be linked easily when matched with the artist name or catalog identifier. Similarly, as the Excerpts database was primarily focused on each unique exhibition entry, the categorization was designed around entries rather than focusing on data identified in other parts of the catalogs, such as introductory pages, in which the majority of the information did not revolve around artworks. After presenting some example categorization formats to the RKD team and fine-tuning them to the project's needs further, the following categories and standards were created:

Categories	**Meaning**	**Reason**	**Example**
Catalogue Identifier	PDF title	Catalogue distinction and database linker	201503502
Keyword Person	Full name of the artist	Matches category title of existing database	A. H. BAKHUIJZEN, Jr.
Artist City	City of artist	Location marker that is often specified in catalogues	te's Hage
Artist Address	Extended Address line if applicable	Additional location marker, split due to being specified in the catalogue less often. Still valuable for insight into artist mobility. Especially for international artists who were more likely to provide full addresses.	unknown
Artist Abbreviations	Catch-all category for exhibition-relevant information, such as membership and competition status	Not all exhibitions used the same abbreviations, and artists sometimes had multiple in the same entry. Catch-all due to non-standardized abbreviations across exhibitions. Done in favour of separate True/False columns for each possible abbreviation.	BL
Entry Number	Listing number of the artwork in the catalogue	Entry distinction	1
Free Title	Title of the artwork	Artwork distinction & matches existing database category	Een Panora
Additional artwork info	Anything listed about artwork that is not part of the title. E.g., Material, quote, ownership, location, etc.	Created as a catch-all due tothe variety of entries and inconsistencies across catalogues	unknown
Asterisk	True/False depending on the presence of the * symbol	While the symbol usually indicates whether an artwork is or is not for sale, this was done in favor of a 'For Sale' category due to the symbol having a variety of meanings throughout the catalogs, which are only found on introductory pages requiring more interpretation work. Keeping track of the asterisk use was particularly relevant for catalogs that do not have a supplementary price list.	False
Amount Type	'Asking Price'	Matches the existing database	Asking Price
Currency	Currency in alphabetical abbreviated label: e.g., HFL	This format matches the existing database, instead of currency symbol or other abbreviations.	unknown
Price	Price of the artwork	Matches existing database + adds price data to artworks	unknown

Full Entry Quote	Complete quote of entry as formatted in the catalogue	Matches existing database & stores data without adjustments	A. H. BAKHUIJZEN , Jr., te 's Hage. B.L. 1 Een Panorama in de omstreken van 's Gravenhage.

Any categories that could not be filled due to missing data in the catalogs were standardized to be indicated with an ‘unknown’ value. Beyond not keeping these cells empty, this choice highlights the challenge of missing data in research and was made to avoid embedding interpretive biases in data that could not be extracted from the catalogs. With these standards set, we could move on to the second phase toward the extraction and structure of the catalogs’ contents.

4. Methodology

4.1. Extraction

In the project, two primary approaches to automate data extraction were explored. The first method involved using **OCR** technologies to extract raw text from scanned documents, followed by **LLMs** to structure the extracted text into a categorized CSV file. The second approach relied entirely on **LLMs** to both extract and structure the data directly.

4.1.1. OCR Engines: Tesseract and Paddle

We engaged with two different open-source OCR Engines to retrieve raw but accurately formatted texts. **Tesseract OCR** was first applied since it is open-access and regarded as one of the most reliable OCR tools due to its ability to handle a wide range of languages, fonts, and its adaptability through custom extensions (Fleischhacker et. al 2024).

As this was a test phase single catalog was processed to evaluate the feasibility and accuracy of the engine. Tesseract OCR does not support reading PDF files, and therefore, the first step involved converting the PDFs into images. To improve recognition accuracy image preprocessing steps were applied, such as denoising and thresholding, and gray scaling. Additionally, to reduce the workload, the first and last pages that do not include relevant content were removed.

After initially experimenting with Tesseract OCR a custom OCR pipeline using **PaddleOCR** with the Dutch (NL) language model was implemented. PaddleOCR is a high-accuracy, open-source OCR engine known for its multilingual support, robust layout handling, and ability to extract fine-grained text features—even from complex or distorted document formats (Sarkar et. al 2024). In our tests with Paddle OCR, the script was applied to multiple catalogs to evaluate its performance across different layouts. The PDF documents were converted into images using the Python library `pdf2image`. Given that the PDFs are scanned printed documents, they can contain handwritten text, bleed-through, and mirror text from the reverse side. Since some catalogs did not require preprocessing, it was applied selectively to catalogs in poor condition to enhance the OCR performance. For consistency reasons, many of the same modules as the Tesseract OCR were reused.

Following this initial setup, the layout structure of the catalogs was addressed. Since most catalogs were saved in two-page spread format (i.e., a single scanned image representing an open book), each scanned PDF page effectively contained two logical pages side by side. To handle this layout correctly, a dedicated function called `split_double_page` was implemented, which automatically divided each image vertically at the midpoint. This allowed to treat the left and right halves of each scan as separate pages preserving the correct reading order and improving the precision of the output.

Each page image was saved locally and then read using OpenCV (cv2) for further processing. OpenCV (cv2) is a tool used for preprocessing the images to enhance their quality and optimize them for text recognition, for example by adjusting contrast, removing noise, or splitting pages. This step helps improve the accuracy of text extraction. The `ocr.ocr` function was then applied to these preprocessed images to perform Optical Character Recognition (OCR), which extracts the actual text content from the image. Each half-page (left and right) was passed through the `ocr.ocr` (Figure 3). The returned results were iterated to extract and concentrate recognized text lines, whereas the outputs were labeled and ordered consistently to reflect natural reading flow. In the last step of the script, the main loop iterates through all PDFs by using a 9-digit PDF identifier (name of the PDF). The extracted raw texts are saved as a .txt file named after the original name of the PDF.

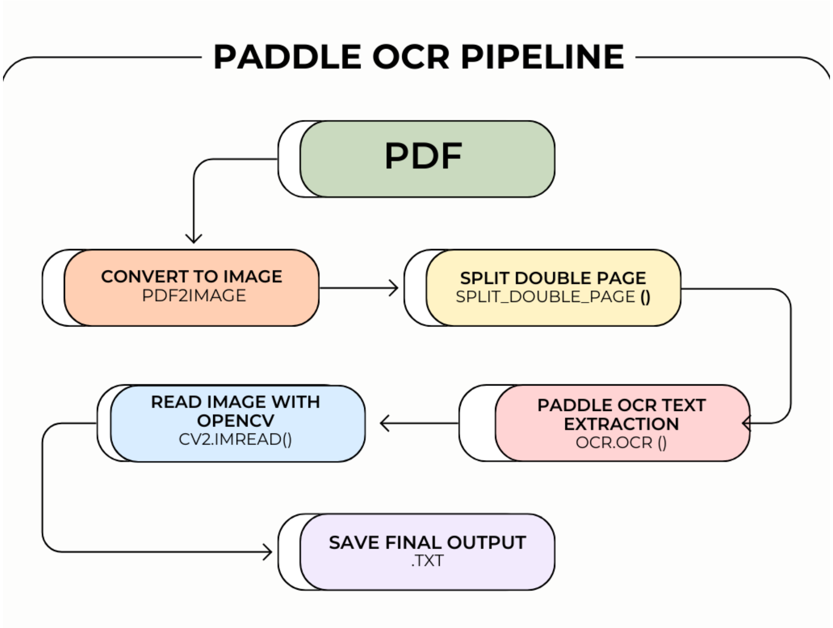


Figure 3.

4.1.2. LLMs

The second method used to extract the data was to use tools by advanced LLM providers: OpenAI and DeepSeek. Different from OCR Engines, LLMs contain semantic understanding, which allows them to identify and interpret information (Solkiran 2025). With the LLM approach, extraction and structuring occurred within a single pipeline as the model was expected to be capable of both identifying the relevant information and organizing it into a structured format in one integrated step.

OpenAI is one of the most well-known companies in artificial intelligence—originally founded in 2015 as a research lab—becoming widely popular in late 2022 with the public launch of the ChatGPT models and ChatGPT, which introduced large language models (LLMs) to a much broader audience (Douglas 2023). These models are designed to understand and generate text, follow instructions, and perform complex tasks such as summarizing, translating, or extracting structured information. (OpenAI 2023).

For this project, OpenAI's GPT-4 model was used through the API. The model allows users to upload text or images (in this case the scanned catalogue pages) together with a prompt specifying the intended task. GPT-4 was particularly useful because it can follow detailed rules, handle messy or inconsistent input, and return structured outputs. This has made it very effective for turning OCR'd text into clear Excel tables.

During this project, Deepseek's models were tested alongside OpenAI. DeepSeek launched their V3 (Chat) and R1 (Reasoner) models in January of 2025, becoming the first open-source models able to rival OpenAI's NLP tasks at a fraction of the cost (OpenAI 2025c; DeepSeek 2025). When called through the API, DeepSeek Chat is better suited for conversational, unstructured, and open-ended tasks, while Reasoner requires more structured input/output goals and goes through an additional process of rethinking and reasoning its own steps (Swiss German University 2025).

4.1.2.1. Open AI

To begin the extraction process with LLMs, we focused on building a pipeline using OpenAI's **GPT-4 Vision API**. The catalogues were first provided as scanned PDF files, which made them unreadable for the Vision API since the model only accepts images. To solve this, each page of the PDF was converted into an image using `pdf2image`. At first, the images were saved in high resolution to make sure all details were visible, especially since some catalogues had faded ink or small font size. This, however, caused issues; when the images were too large, either in resolution or in content, the API struggled to return a complete response. Sometimes the model would stop halfway through or return an error. This happened because the model has a limit on how much data—in terms of image size and internal token count—it can handle in one go.

To fix this, the images were resized to around 200-300 DPI, which kept the text clear while reducing the amount of information per image. They were also converted to grayscale to shrink the file size even further without making the text harder to read. After this, each image was encoded in `base64`, a way of converting files like images into plain text using only letters, numbers, and symbols. In our case, it has been used to turn each page image into a long string of text that the model could read. This allowed to send the image directly through the API request without needing to upload the file somewhere else. While `base64` does increase the size of the image data, it makes the process more flexible and easier to automate.

4.1.2.2. DeepSeek

When tasking the DeepSeek models through the API to extract the text, significant changes had to be made. The chat model only allowed a very limited amount of image processing making it incompatible with the large volume of images the existing pipeline required to run. Instead, rather than converting the PDFs first, they were "fed" to the API as is, without any preprocessing steps applied. This was feasible by using `PyMuPDF`'s `fitz` module to allow the model to read the file and extract the text from the PDFs in a readable format directly. It was then prompted to read, extract, and structure the text according to the defined categorizations.

4.2. Structuring

4.2.1. Transforming OCR Output into Structured Data Using LLMs

After extracting raw text from the scanned catalogues using OCR engines, the next step was to convert the raw text into a clean, tabular Excel file. To carry out the structuring phase, both the **DeepSeek Chat Model** and the **OpenAI GPT-4 Turbo model** were utilized. When structuring the OCR output, there was no notable difference in the pipeline between the two LLMs; both followed a similar process for interpreting and formatting the extracted text.

The OCR output was initially stored as .txt files, containing blocks of text representing catalogue entries. To assist the LLMs, extensive prompts were provided to explain where the information is located and how it should be stored. This rule-based parsing set consisted of examples on how to save the artist's name (extract the first name first and then the surname), how to handle abbreviations, and how to store missing information (if no information is available, store as "unknown"). Additionally, a rule was created to ensure that the price list aligned with the correct entry number and is saved accordingly. Since the price list is usually found on a separate page, it was important to emphasize in the prompt exactly how it should be stored.

Rules for handling ambiguous catalog notations like "idem" or "dezelfde," and specifying how to structure them were also included in the prompt. These are Dutch shorthand words that mean "same as above," and they are often used to refer to the previous artist's name or location.

Once the model returned the structured output, the response was parsed using Python's `pandas` and `StringIO` libraries to create a dataframe for each segment, which was then merged into a single table. Finally, the complete dataset was exported as an Excel file using `openpyxl`. Throughout this process, the script maintained a high level of detail, preserving the original entry order, handling inconsistent formatting, and filtering out elements such as irrelevant abbreviations—ensuring that the final output remained both accurate and usable.

4.2.2. Excel Table Generation from API Output Using LLMs

4.2.2.1. Open AI

Once the information was extracted from the catalogs using the GPT-4 Vision API, it needed to be organized into a structured format. The initial intention was to have the model return everything in JSON format, a standard way of storing structured data. The model often returned broken or messy JSON—missing commas, unescaped characters, or wrong formatting. Even when given correct examples, the model would sometimes produce output that was unreadable for the parser. Consequently, the results were ultimately stored in an Excel file format instead.

A crucial part of the structuring process was prompt engineering. Because the model's output depends heavily on instructions, prompts were crafted to extract and structure specific fields from each image: the artist's name, city or address, entry number, artwork title, parenthetical notes, asterisks, and known abbreviations (e.g., GL, BM). These prompts were based on a custom rule document and dynamically adjusted depending on the page layout. This flexibility allowed for a more responsive and adaptive extraction process.

During the process, several edge cases appeared that required special attention. One of the most common issues encountered in the structuring phase of OCR extracted text was with entries that used the words "idem," "dito," or "dezelfde." At first, the model copied these words directly into the output, which made the dataset incomplete. To fix this, the prompt was updated to inform the model that when one of these words appears it should reuse the last extracted artist name and location. This worked in many cases, but to ensure, an extra Python function was added after the extraction step. The function looked for those words and automatically replaced them with the correct information from the row above. There were also entries where the artist's name was split across lines or where the name included extra information like titles or unnecessary abbreviations. For example, abbreviations like "GL," "M," or "BM" were sometimes added after the name or location, but they were not useful for the dataset and were later removed.

To structure prices, which often appear in the catalogues in a separate section labeled "Prijzenlijst", a different prompt was used. The format usually includes an entry number, a currency symbol (like "f" for gilders or "m" for marks), and the price. Initially, it was considered getting the model to extract prices while processing the artwork entries, but this did not return the expected results, since the price lists are often formatted differently and are sometimes not directly linked to the artworks. Instead, these pages were processed independently, prompting the model to extract only entry numbers and price values. These were later matched back to the artworks using entry numbers as a shared reference point.

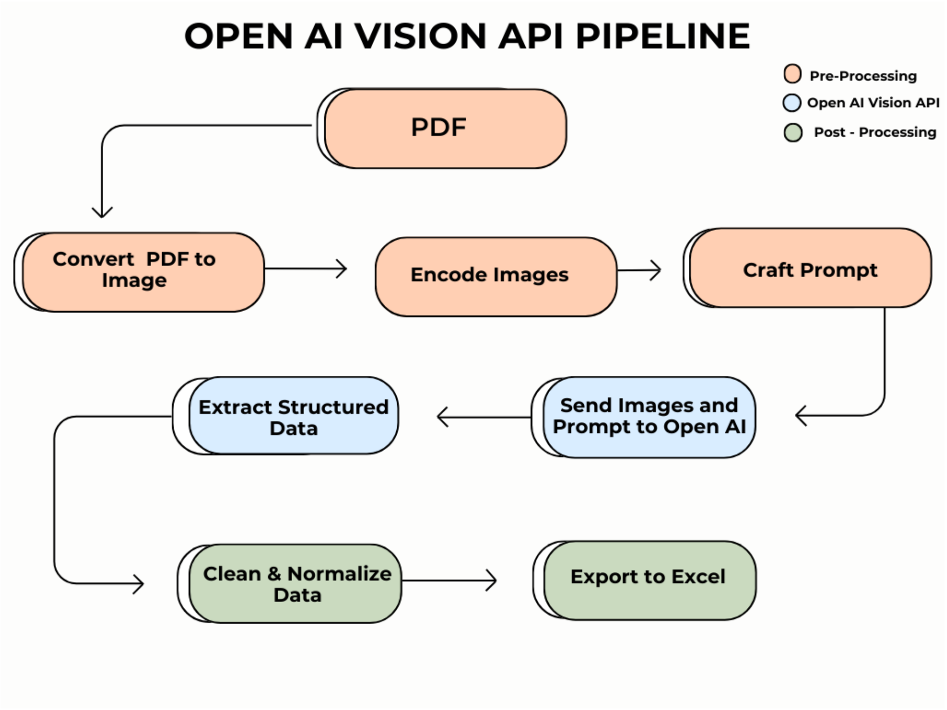


Figure 4.

4.2.2.2. DeepSeek

The first method of extracting raw text through OCR and structuring it via the DeepSeek API remained largely unchanged, due to DeepSeek's reliance on text-based input. However, in this case, the text was extracted directly from the PDFs. Prompting the DeepSeek API to structure the text according to the defined categorizations and edge cases mentioned above was done in a slightly different manner. While all core aims and functions remained the same, there were some formatting variations in the DeepSeek prompt. For example, the prompt was formatted in a more instructive manner, incorporating more specific conditions and rules for each category. This also included a line instructing the model to correct spelling mistakes that may have occurred during the text extraction. In contrast to the OpenAI pipeline, prompting was not divided between separate instructions for the entries and the price list, nor was an additional Python function implemented to address edge cases such as "idem". All other steps remained largely unchanged, aside from minor adjustments for API compatibility.

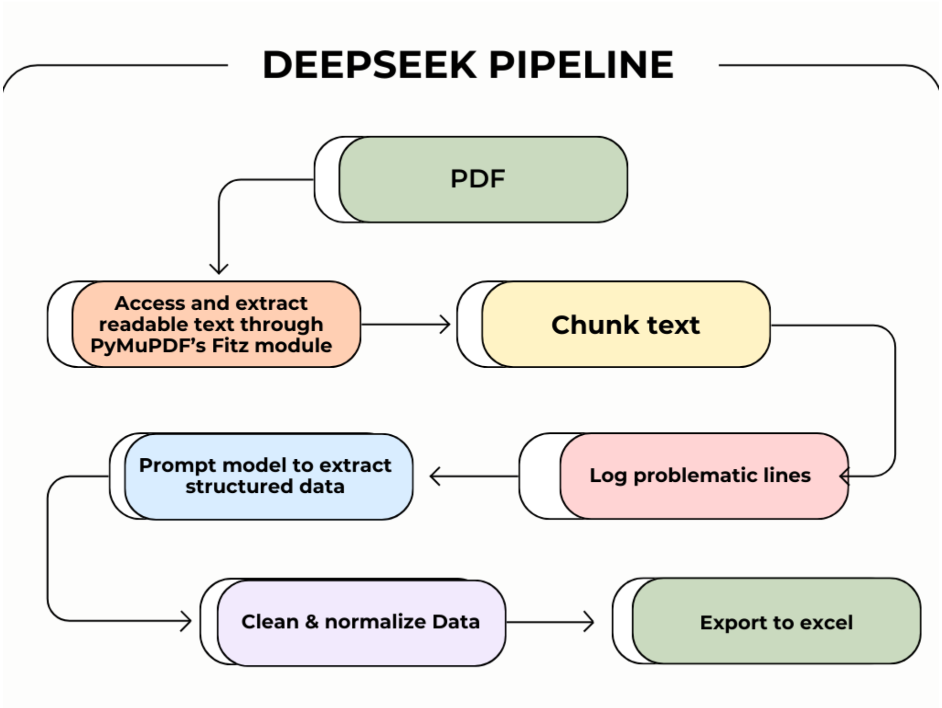

DEEPSEEK PIPELINE
PDF
Access and extract readable text through PyMuPDF's Fitz module
Chunk text
Log problematic lines
Prompt model to extract structured data
Clean & normalize Data
Export to excel

Figure 5.

5. Results

5.1. OCR & LLM Processing Pipeline

Using OCR engines in extracting the raw text gave mixed outcomes. At first, Tesseract OCR was used without preprocessing. The results consisted of a lot of typo errors. For example, the city name “Amsterdam” was misspelt as “Amstederpam”. Similar issues appeared across other catalogues: “Rotterdam” was read as “Rotteldam,” and “Utrecht” occasionally appeared as “Uiricht.” In order to prevent the misspellings, an image preprocessing step was performed. However, the adjustments had little to no effect on improving accuracy. The same errors persisted in the output, suggesting that Tesseract struggled with the specific fonts and quality of these specific historical documents’ scans.

In contrast, the results improved significantly when experimenting with PaddleOCR. While still producing raw, unstructured output, the visual layout of the entries was preserved more accurately, making the text easier to work with in the next step of the pipeline (Figure 6). Despite the more accurate extractions in general, price extraction remained a major challenge. As previously noted, price lists are not embedded within the main text body but instead listed separately, sometimes pages apart from the artwork entries, which makes it difficult for the

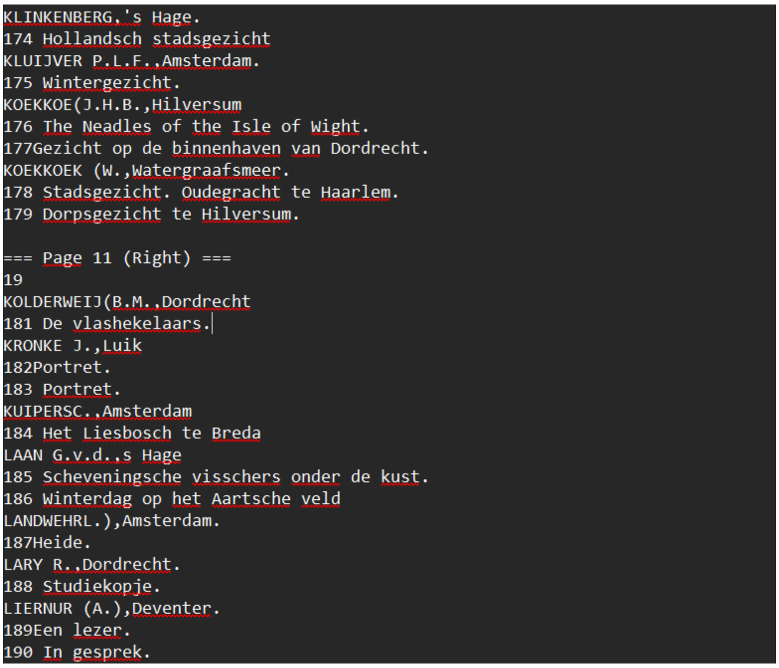
OCR to capture price information from the raw text. Additionally, price lists have a complex layout, consisting of two, three ,and sometimes four columns and therefore, OCR has often failed to link prices with the correct artwork entries. Another significant challenge during extraction was the confusion between the lowercase letter “1,” the number “1,” and the uppercase letter “I.” Misspelling these characters led to inaccuracies in the raw text—particularly in names, titles, and catalogue numbers—which in turn affected the reliability of the structured data.

In addition, some catalogues contained handwritten annotations, stains, or blurred characters. As an OCR engine, PaddleOCR attempts to extract all visible text—including handwritten content—which introduced further inconsistencies in the output. Preprocessing steps aimed at filtering out handwritten elements did not significantly improve performance in these cases, and recognition errors persisted.

In the structuring phase**,** outputs from PaddleOCR were tested, focusing primarily on catalogues with clearer layouts and more consistent formatting. For this stage, both DeepSeek Chat and OpenAI’s GPT-4 Turbo were applied to transform the raw text into a structured Excel format. The models were particularly successful at extracting and organizing ‘Keyword Person’, ‘Artist City’, ‘Full Entry Quote’, and ‘Additional Artwork Information’. When comparing the two models, OpenAI’s GPT-4 Turbo outperformed DeepSeek Chat in terms of both speed and accuracy. The outputs generated by GPT were typically more thorough, especially in handling longer entries or catalogues with slight layout variations.

However, structuring the raw text revealed significant limitations and challenges. One of the main issues was implementing the “dezelfde” and “idem” rules. When an artist was listed in multiple artworks consecutively, the model often failed to retain and assign the artist’s name to each entry individually. As a result, the ‘Keyword Person’ column sometimes contained missing values.

A further complication was matching the prices with entry numbers. Although the model occasionally matched the prices correctly, the majority of the entries were wrong or missing. The problem intensified in cases where artworks were missing from the price list, resulting in misaligned rows and inconsistencies in the structured output. Furthermore, the structuring of currency values proved problematic, largely due to irregularities in the raw OCR output. In the raw text output, the OCR engine frequently failed to position currency symbols adjacent to their corresponding prices. As a result, the LLMs often did not recognize them as related elements, leading to inconsistent or missing currency data. Another issue occurred in storing abbreviations, as the respective column was often filled inaccurately or with the "unknown" value.



```
KLINKENBERG.'s Hage.
174 Hollandsch stadsgezicht
KLUIJVER P.L.F..Amsterdam.
175 Wintergezicht.
KOEKKOE(J.H.B.,Hilversum
176 The Neadles of the Isle of Wight.
177Gezicht op de binnenhaven van Dordrecht.
KOEKKOEK (W.,Watergraafsmeer.
178 Stadsgezicht. Oudegracht te Haarlem.
179 Dorpsgezicht te Hilversum.

=== Page 11 (Right) ===
19
KOLDERWEIJ(B.M..Dordrecht
181 De vlashekelaars.
KRONKE J.,Luik
182Portret.
183 Portret.
KUIPERSC..Amsterdam
184 Het Liesbosch te Breda
LAAN G.v.d..s Hage
185 Scheveningsche visschers onder de kust.
186 Winterdag op het Aartsche veld
LANDWEHRL.),Amsterdam.
187Heide.
LARY R..Dordrecht.
188 Studiekopje.
LIERNUR (A.),Deventer.
189Een lezer.
190 In gesprek.
```

Figure 6. Raw Text Output (PaddleOCR)

5.2. LLM Processing Pipeline

The GPT-4 Vision proved to be a highly capable tool for structuring data from scanned art catalogues, particularly because of its ability to understand both layout and textual context. For instance, the model was often able to correctly associate multiple artwork entries listed beneath a single artist name, even when the name was only written once and followed by several indented titles. In these cases, the model maintained the correct artist for each entry without duplicating or dropping information. It also handled notes that appeared in parentheses or as indented lines, such as "(on loan from the artist)" or "(painted during his time in Paris)" and placed them under 'Additional Artwork Information' rather than misinterpreting them as part of the title.

Another strength was how it dealt with edge cases like "idem", "dito", or "dezelfde". After prompt adjustments, the model began substituting these correctly with the previous row's values rather than copying the word directly into the dataset, which had been a recurring issue from early on. Entries that were split across lines due to formatting—such as when titles wrapped awkwardly or when abbreviations were placed on a new line—were also handled correctly. It was usually able to combine these into a single, clean title. In some cases, it even ignored decorative or stylistic elements—like catalog headers or exhibition section labels—that were not relevant to the entry itself, improving the clarity of the output.

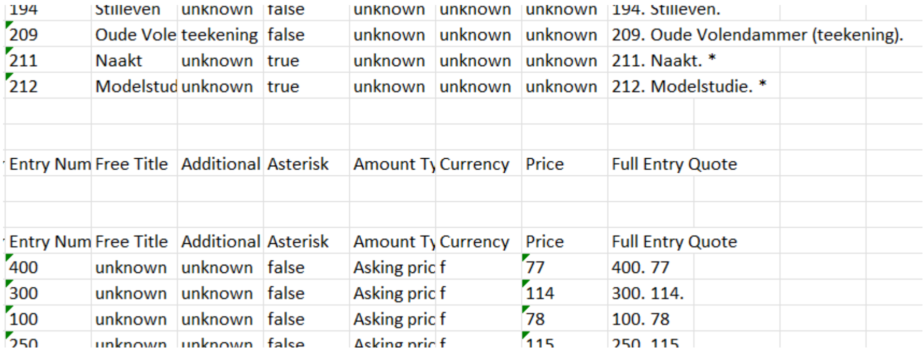
Additionally, GPT-4 Vision was successful in identifying the handwritten texts in the catalogs, often consisting of price notes or comments added by visitors. Unlike OCR engines, it was able to identify and exclude these handwritten elements, preventing unrelated information from being included in the structured output.

However, some issues remained. One recurring problem was in the 'Artist Abbreviations' column; the model sometimes confused name initials or infixes with abbreviations like "GL", "BM", or "M", resulting in either incorrect entries or blank fields. Another example was the handling of layout differences across catalogs. In some cases, the title of the artwork is printed next to the artist's name rather than below it. Although the model has proven to be generally adaptive, its accuracy dropped slightly when the catalog structure differed significantly from the one used during prompt training. Similarly, formatting inconsistencies such as unexpected use of capitalization, punctuation, or spacing could affect the way the model parsed names and addresses, occasionally leading to misalignment in the output columns.

While the OpenAI pipeline did not fully match the original set of categorization columns, it was still able to extract many important fields with consistent and usable results. The most reliable outputs were found in the 'Keyword Person', 'Artist City', 'Entry Number', 'Free Title', 'Additional Artwork Information', 'Asterisk', and 'Full Entry Quote' columns. In most cases, the model correctly linked multiple artworks to the right artist—even when the artist's name appeared only once at the top of a list. Titles of artworks were placed clearly in the 'Free Title' column, and extra information—such as text in parentheses or notes about the piece—was usually saved in the 'Additional Artwork Information' field. Asterisks were correctly detected and stored as true or false values. The 'Full Entry Quote' column was also handled quite well, with most entries returned in full and with minimal formatting issues. At the same time, other fields that were planned, such as 'Artist Abbreviations', 'Artist Address', and 'Amount Type', were less reliable as the model often confused artist initials with exhibition-related abbreviations or missed the abbreviations entirely. Artist addresses are rarely clear in the catalogues and often merged with city names, making them difficult to extract accurately. The 'Amount Type' column—which was meant to default to 'Asking Price'—added little value because that kind of information was not mentioned in most catalogues and repeated the same text for every row. Ultimately, these fields were removed or simplified in the final Excel files to avoid confusion and keep the output cleaner and accurate.

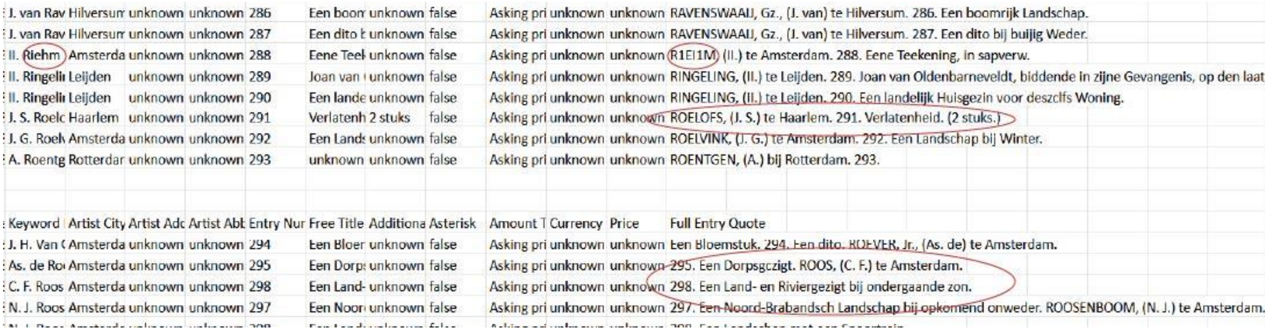
Prompting with DeepSeek, on the other hand, required more explicit instructions per column to work as intended but was otherwise quite intuitive; it did not need any example format in the prompt, and was able to match the stated extraction and structuring rules to most of the right columns sufficiently by explaining the purpose of each output column. This included instructing it to not only extract names but also standardize their format to title-first name-infix-surname. It was also able to grasp the meaning of "deselfde/dezelfde" and structure accordingly without prompting it or adding other functions to do so, even though this was still added to the final prompt to standardize the rule across catalogs and to include less self-explanatory terms with the same meaning, e.g., "ditto". These apply to both the Chat Model and the Reasoner Model. Additionally, they were sometimes able to correctly extract and structure the entries in the 'Full Entry Quote' column with minimum cleaning, however, this result was not very consistent (Figure 8). In some cases, the LLM fixed spelling mistakes in the other columns of a row but kept the full quote as it was extracted from the PDF. As such, instances occurred where "ll" was output rather than "11" or "I5" rather than "B" specifically in this column. Further, due to token limits, to which we will refer in a later section, a segmentation function was added to the pipeline to parse through entire catalogues without cutting off, e.g., 50 entries in. Due to the splitting, the 'Full Entry Quote' column could behave slightly differently depending on what part was being processed, leading to small inconsistencies, e.g., writing the entries but not the artist information or the opposite.

Regarding other challenges, some columns were not able to be extracted and structured sufficiently with this pipeline, such as the 'Price', 'Currency', and 'Artists Abbreviations'. At best, it was able to extract prices separately but did not match them to existing entry lines (Figure 7). These fields were often left empty, whereas the extraction of 'Additional Artwork Information' proved inconsistent, varying considerably from one entry to another.



194	Stilleven	unknown	false	unknown	unknown	unknown	194. Stilleven.
209	Oude Vole	teekening	false	unknown	unknown	unknown	209. Oude Volendammer (teekening).
211	Naakt	unknown	true	unknown	unknown	unknown	211. Naakt. *
212	Modelstud	unknown	true	unknown	unknown	unknown	212. Modelstudie. *
Entry Num	Free Title	Additional	Asterisk	Amount Ty	Currency	Price	Full Entry Quote
Entry Num	Free Title	Additional	Asterisk	Amount Ty	Currency	Price	Full Entry Quote
400	unknown	unknown	false	Asking pric	f	77	400. 77
300	unknown	unknown	false	Asking pric	f	114	300. 114.
100	unknown	unknown	false	Asking pric	f	78	100. 78
250	unknown	unknown	false	Asking pric	f	115	250. 115

Figure 7. Deepseek-Chat price results.



J. van Rav	Hilversum	unknown	unknown	286	Een boom	unknown	false	Asking pri	unknown	unknown	RAVENSWAAIJ, Gz., (J. van) te Hilversum. 286. Een boomrijk Landschap.
J. van Rav	Hilversum	unknown	unknown	287	Een dito b	unknown	false	Asking pri	unknown	unknown	RAVENSWAAIJ, Gz., (J. van) te Hilversum. 287. Een dito bij buijig Weder.
II. Riehm	Amsterda	unknown	unknown	288	Eene Teek	unknown	false	Asking pri	unknown	unknown	R1EI1M (II.) te Amsterdam. 288. Eene Teekening, in sapverw.
II. Ringeli	Leijden	unknown	unknown	289	Joan van (	unknown	false	Asking pri	unknown	unknown	RINGELING, (II.) te Leijden. 289. Joan van Oldenbarneveldt, biddende in zijne Gevangenis, op den laat
II. Ringeli	Leijden	unknown	unknown	290	Een lande	unknown	false	Asking pri	unknown	unknown	RINGELING, (II.) te Leijden. 290. Een landelijk Huisgezin voor deszclfs Woning.
J. S. Roelc	Haarlem	unknown	unknown	291	Verlatenh	2 stuks	false	Asking pri	unknown	unknown	ROELOFS, (J. S.) te Haarlem. 291. Verlatenheid. (2 stuks.)
J. G. Roelv	Amsterda	unknown	unknown	292	Een Lands	unknown	false	Asking pri	unknown	unknown	ROELVINK, (J. G.) te Amsterdam. 292. Een Landschap bij Winter.
A. Roentg	Rotterdar	unknown	unknown	293	unknown	unknown	false	Asking pri	unknown	unknown	ROENTGEN, (A.) bij Rotterdam. 293.
Keyword	Artist City	Artist Adc	Artist Abl	Entry Nur	Free Title	Additiona	Asterisk	Amount T	Currency	Price	Full Entry Quote
J. H. Van (	Amsterda	unknown	unknown	294	Een Bloer	unknown	false	Asking pri	unknown	unknown	Een Bloemstuk. 294. Een dito. ROEVER, Jr., (As. de) te Amsterdam.
As. de Ro	Amsterda	unknown	unknown	295	Een Dorps	unknown	false	Asking pri	unknown	unknown	295. Een Dorpsgczigt. ROOS, (C. F.) te Amsterdam.
C. F. Roos	Amsterda	unknown	unknown	298	Een Land-	unknown	false	Asking pri	unknown	unknown	298. Een Land- en Riviergezigt bij ondergaande zon.
N. J. Roos	Amsterda	unknown	unknown	297	Een Noor	unknown	false	Asking pri	unknown	unknown	297. Een Noord-Brabandsch Landschap bij opkomend onweder. ROOSENBOOM, (N. J.) te Amsterdam.

Figure 8. DeepSeek-Chat segmented results.

Furthermore, despite performing structuring tasks effectively, the pipeline often failed to extract entire catalogues, frequently skipping entries even with the chunking function in place. This issue was likely attributable to the way both Python and the model parsed the text source. This problem was more evident when the text input was only the PDF file compared to an OCRed text file. Another limitation was the Reasoner model's tendency to overcorrect entries; instead of simply addressing spelling errors, it often introduced significant and consistent alterations to titles. These findings underscore minor compatibility issues between the current capabilities of the models and the scale and complexity of the data they were required to process.

6. Discussion/ Limitations

In light of the results, from a practical perspective, the OpenAI model has some downsides. It is only available to paying users and requires access to the GPT-4 tier, which can make it inaccessible for researchers with limited or no budget. Additionally, the cost becomes a significant factor when handling large quantities of images, given that each API call is billed individually. On top of that, the model is still being updated regularly by the company, and those updates can change how it behaves. For example, sometimes prompts that worked one week would give different results the next, which made the workflow unpredictable at times. Despite these challenges, the flexibility and power of the GPT-4 Vision made it possible to build a semi-automated system that could extract structured, usable data from historical art exhibition catalogs, something that would have taken much longer to do manually.

In comparison, while the DeepSeek API incurred lower costs and was easier to access in theory by not requiring a credit card to top up the balance, the calls were limited to its Chat and Reasoner models. This means any extracting or structuring was mainly applicable to text input compared to OpenAI, which also currently embeds extensive vision capabilities in its models.

Furthermore, the workflow using DeepSeek's API was unpredictable at times due to occasional top-up limits that the company would impose as a result of its own more limited hosting resources. Additionally, DeepSeek generally had a lower output token maximum or context window than OpenAI (OpenAI 2025a, 2025b; DeepSeek 2025). This required additional segmentation, leading to inconsistencies during the extraction and structuring process. In the end, DeepSeek and OpenAI outputs had some very apparent differences in their functions and outputs.

That said, the most pronounced distinctions between the two models lie in their cost-effectiveness and processing efficiency. While DeepSeek's Chat model produced results that were comparable to OpenAI's in many cases, its main advantage was its lower cost. However, when it comes to efficiency, OpenAI performed better. DeepSeek required additional setup and preprocessing steps to reach similar outcomes. This was evident in the integration of price lists, and when OCR-extracted text input yielded better results with DeepSeek than its direct extraction. However, price list extraction remained problematic when using OCR alone. While the OpenAI results require some extra steps, these mainly pertain to less intensive cleaning during post-processing.

Based on this analysis, we conclude that the OpenAI pipeline is more productive and compatible with the aims of this project and should be prioritised moving forward. However, despite this prioritization, the experimentation with both LLMs highlighted some promising capabilities and future considerations.

Beyond the aforementioned results, this research showcases new possibilities within the digital art history field. We argue that it is possible to automate the art exhibition catalog extraction and structuring process to a significant extent using computational methods, specifically utilizing newly emerging AI technologies. This highlights the potential of a new workflow toward open data (Figure 9), one that reduces the need for intensive manual labor while accelerating the acquisition, sharing, and standardization of art historical knowledge, reflecting Swartz's (2019) positioning of the digital as a tool for expansion in the discipline.

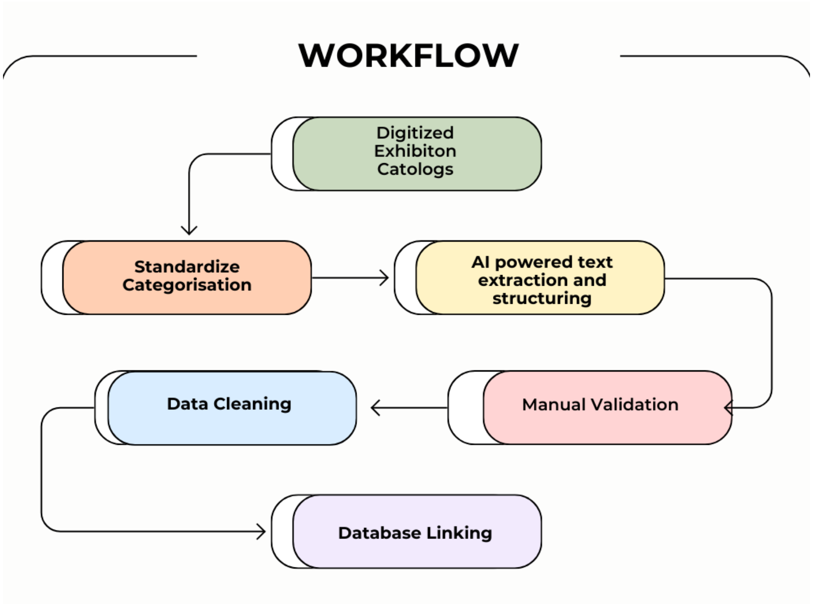

WORKFLOW
Digitized Exhibiton Catologs
Standardize Categorisation
AI powered text extraction and structuring
Manual Validation
Data Cleaning
Database Linking

Figure 9.

Moreover, we attempt to showcase a largely non-intrusive text-based method for DAH, affirming that digital methods do not have to replace human interpretation and historical analysis. We argue that text-based computational methods have a valuable role in preprocessing workflows that operate independently of interpretive text analysis. This minimizes the risks proposed by Bishop (2018) in undermining the depth of art historical interpretation in research.

Despite these promising results, further development of the proposed methods—particularly in the form of data cleaning and manual validation by experts in the field—is necessary before outputs from this tool can be considered suitable for publication. While LLMs are capable of correcting OCR errors, they are not exempt from bias and can introduce inaccuracies that require careful monitoring (Thomas et al. 2024). Even though our findings indicate that they can outperform other spell-correction tools (Do et al. 2025), they can still be prone to overcorrecting and need to be monitored to avoid creating and preserving faulty data. Ultimately, elements such as prompt quality can result in LLM hallucinations, also leading to the generation of nonsensical and false outputs that require closer inspection.

Further, in our findings, a relationship was observed between the performance of the models and deviations in the catalogs; the model was able to perform better with catalogs that only had small differences in their formatting, compared to catalogs with more robust formatting variations. Therefore, while LLMs demonstrate promising capabilities in semantic understanding, their outputs are not flawless, and some linguistic and formatting discrepancies persist more prominently than others. As such, the extent of format deviation remained a challenge for the models, leading to slight performance differences or inconsistencies, thereby highlighting the need for expert supervision and validation.

Throughout the process of this research, the interdependencies of using the models were also encountered. Not only were the models steered to behave according to our goals, but the affordances of each model would also significantly affect the workflow in return. One notable case was prompting: although both models were tasked with the same goals, the structure and style of their prompts differed significantly. Testing a model was rarely as simple as switching the API key and loading the correct model; in practice, prompts often had to be carefully fine-tuned to suit the specific model in use. While some instructions were understood by both models, overall DeepSeek required more instructive language, while OpenAI was more responsive to succinct labelling and example inputs, showcasing that LLMs do not all interpret or process natural language the same way, which is an important factor to consider when deciding which model(s) to use in a workflow.

Similarly, the cost of the models affected the workflow more than anticipated. For example, a low-cost model like DeepSeek was more often used to experiment with multiple prompts and larger batches of test catalogs. This was especially relevant when re-running the pipeline based on minimal fine-tuned adjustments to the prompt to check how it would perform, making experimentation with the model less restrictive. Running it through an OpenAI model instead often came with more cost consciousness. This was revealed partly in the prompting, such as minimizing the input size of the prompt along with adjusting it less extensively throughout the process, but more so in testing the catalogs. The OpenAI prompts were typically tested on batches of no more than 3-6 catalogs, limiting the assessment of the model's performance at scale. In this case, costs could be a barrier to experimentation. Naturally, however, the influence of costs on the workflow depends on the project. In our case, due to the experimental nature of the research, we aimed at keeping the costs as low as possible, leading to these workflow differences. In a project where cost is less of a constraint, such limitations are likely to be less disruptive to the overall workflow.

Additionally, in our project, the collaboration with the RKD significantly aided the quality of the process and the results. Throughout the exploration and analysis of the catalogs, we maintained an ongoing dialogue with a team of art historians and computer scientists. This multidisciplinary approach increased our understanding of the corpus and the structure of their existing databases, making the standardization process smoother. Similarly, any results encountered were often discussed with the team, which brought forth new ideas for fine-tuning prompts or validation methods, highlighting the benefits of assembling multidisciplinary teams for this kind of research.

While there is significant potential for multimethod pipelines in the digital humanities—by exploring new technological advancements and combining them with already explored tools such as OCR—these methods should be approached by staying mindful of the limitations of biased, deviating and hallucinatory outputs paired with considerations regarding tool accessibility, cost, efficiency and expert supervision.

7. Recommendations for the RKD and Future Researchers

Based on the findings of this project, several recommendations can be made to improve the use of OCR and LLM-based workflows for historical art exhibition catalogues. These suggestions aim to increase accuracy, reduce manual cleanup, and support long-term scalability while remaining mindful of accessibility and available resources.

Firstly, in terms of tool selection, we recommend using OpenAI's GPT-4 Vision as the primary model for structuring catalogue data. Its ability to follow complex instructions and understand layout-based cues made it particularly effective when working with varied catalogue formats and entries that required nuanced interpretation. While the model comes with access and cost limitations, its performance in linking artists to artworks, recognizing patterns like "idem," and organizing multi-line entries outweighed those of other tested tools.

However, for resource-constrained projects or less complex formatting needs, DeepSeek offers a valuable alternative as it performed well, especially in structuring pre-extracted text and was able to handle many of the project's core tasks at a much lower cost with some occasional inconsistencies. In its current state, DeepSeek's strengths lie in text processing and summarization. Therefore, its capabilities may be more compatible with extracting summary points from larger texts, requiring less extensive prompting. In the case of the catalogues, it may be more effective to use the model exclusively for structuring tasks by first providing a pre-extracted text document, thereby mitigating issues related to skipped entries. When it comes to reconciling the price lists with the catalogs, it would likely be more fruitful to separate these during prompting completely and to append the separate dataset of prices to the catalog data after the AI-powered extraction and structuring process. The API should also be revisited when the platform rolls out a more extensive image processing model, and remains a valuable alternative to increase competition within the AI market.

Regarding OCR, PaddleOCR using the Dutch language model provided more accurate and visually coherent results than Tesseract, especially when working with older typefaces or degraded pages. To ensure more consistent OCR results, it is recommended to apply standard preprocessing steps such as grayscale conversion, noise reduction, and careful image splitting, especially when dealing with two-page spreads. Additionally, OCR and LLM combination is recommended for the extractions involving a smaller number of categories.

When optimizing extraction accuracy is concerned, prompt engineering plays a key role throughout the project and should be treated as a flexible, evolving component of the pipeline. Prompts should clearly instruct the model on how to handle repeating values such as “idem” or “dito,” how to extract and format full artist names, how to distinguish between true abbreviations and name parts, and how to match price list entries using the artwork’s entry number rather than proximity on the page. Maintaining a well-structured prompt template that can be slightly adjusted depending on the catalogue layout can help increase output consistency and save time across large-scale processing.

Additionally, to improve results, it would be helpful to standardize input PDFs where possible in future digitization efforts. This could include scanning pages individually rather than in double-page spreads, ensuring scans are clear and high-resolution, and avoiding inclusion of marginalia or handwritten notes when not relevant. In catalogues, however, where re-digitization is not feasible, applying consistent preprocessing (grayscale conversion, denoising, cropping) before OCR is strongly encouraged. Similarly, separating price list pages from entry sections during scanning or editing would reduce confusion and improve matching accuracy.

While automation by using GPT-4 Vision can significantly reduce manual workload, it is not free from errors. Therefore, we recommend applying a series of post-processing rules to improve the consistency and usability of the extracted output.

One common issue in the tables is the presence of square brackets around values, for example, [Amsterdam]. To remove these brackets, Excel’s Visual Basic for Applications (VBA) can be used by applying a simple rule to delete the brackets while preserving the text. Furthermore, entries in the ‘Free Title*’*, ‘Additional Artwork Information’, and ‘Full Entry Quote’ columns should be cleaned to remove repeated line breaks and unnecessary punctuation.

As discussed in the results part, the most frequent errors occurred within the price and currency columns. Therefore, we recommend that a domain expert manually reviews and corrects the price column to ensure the dataset is clean and reliable. In some cases, the price column contains the currency value appended to the numbers. Without a manual check, the currency value can be quickly erased from the column by using a VBA script that removes any surrounding text or symbols and stores merely numeric digits.

Additionally, currency columns require manual cleaning and correction similar to price columns. Since there is no consistent value for currencies in the catalogs, the currency symbols should be standardized in line with the RKD's database format (e.g., MRK for mark and HFL for gulden). Also, the quotation mark symbol (") should be replaced with the value from the previous row when used to indicate repetition. These corrections can be automated using Excel's VBA rules.

Ambiguity and missing data are common challenges. As a rule of thumb, we recommend continuing to use the value "unknown" in fields that cannot be reliably filled, rather than leaving them blank or guessing based on context. This avoids interpretive bias and helps future researchers see where data gaps exist. In cases where "unknown" may be inconvenient—like English-language documents in which "unknown" can signify the name of the artwork—similar placeholders like "untitled" or "no title given" may be more fitting. The key is to remain consistent and transparent about what is known and what remains uncertain.

For recurring shorthand like "idem" or missing currency symbols, handling should be done both within the prompt and during post-processing. Adding small Python functions that detect these cases and reuse or fill in data from previous rows proved effective in improving dataset quality.

Additionally, during batch processing, it is helpful to keep a log of all errors, skipped entries, or inconsistencies. This not only helps with troubleshooting but can also reveal patterns in where the pipeline requires refining.

Looking ahead, we suggest exploring the potential of fine-tuning a smaller language model on a set of manually labeled catalogues. This could improve precision in areas that LLMs still struggle with, such as abbreviation parsing or distinguishing between similar formatting structures. Moreover, if the goal is to make this workflow more accessible to non-technical users, it may be valuable to package it into a lightweight tool or web interface. This could allow users to upload a catalogue and receive a structured Excel or XML output without needing to work directly in a coding environment.

As a potential next step, a crucial area of exploration would be the development of an AI-assisted correction model that can compare the catalog PDFs with the Excel table output to identify mismatches or errors automatically. Since AI models are evolving rapidly, integrating them for post-extraction quality control could be a valuable addition to the pipeline. This would not only improve efficiency but also enhance the reliability of the final dataset, contributing meaningfully to future automation efforts at RKD and beyond. By combining careful prompt design, consistent preprocessing, and the strengths of modern AI and LLMs, cultural heritage institutions can move closer to large-scale, reliable, and (semi)-automated structuring of historical records.

8. Conclusion

This project demonstrated the feasibility of automating the extraction and structuring of historical documents using OCR and LLMs, with a focus on the 19th century Living Masters art exhibition catalogues at the RKD. By comparing OCR-based and fully LLM-driven pipelines, we showed that tools like PaddleOCR and GPT-4 Vision can significantly reduce manual labor while preserving a level of accuracy. Although challenges remain, such as handling inconsistent layouts, abbreviations, and price associations, the results highlight the value of integrating AI into digital art history workflows. Beyond technical experimentation, this work also contributes a replicable methodology that bridges computational innovation and archival scholarship, offering a scalable path forward to GLAM institutions seeking to unlock unstructured historical text collections.

References

1. Abdellaif, Osama Hosam, Abdelrahman Nader Hassan, and Ali Hamdi. 2024. “LMRPA: Large Language Model-Driven Efficient Robotic Process Automation for OCR.” *arXiv* preprint arXiv:2412.18063. https://arxiv.org/abs/2412.18063.
2. Amangazykyzy, Moldir, Aigerim Gilea, Aubakirova Karlygash, Abisheva Nurziya, and Kulanova Sandygash. 2025. “Epistemological Transformation of the Paradigm of Literary Studies in the Context of the Integration of Digital Humanities Methods.” *Forum for Linguistic Studies* 7 (4): 166–176. https://doi.org/10.30564/fls.v7i4.8619.
3. Baca, Murtha, Anne Helmreich, and Melissa Gill. 2019. “Digital Art History.” *Visual Resources* 35 (1–2): 1–5. https://doi.org/10.1080/01973762.2019.1556887
4. Bago, Petra, and Nikola Ljubešić. 2015. “Using Machine Learning for Language and Structure Annotation in an 18th Century Dictionary.” In *Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age*, 427–442. https://elex.link/elex2015/proceedings/eLex_2015_28_Bago+Ljubesic.pdf
5. Bode, Katherine, and Charlotte Bradley. 2024. “Computational Literary Studies and AI.” In *The Routledge Handbook of AI and Literature*, edited by Will Slocombe and Genevieve Liveley, 9. 1st ed. New York: Routledge. https://doi.org/10.4324/9781003255789.
6. Bishop, Claire. 2018. “Against Digital Art History.” *International Journal for Digital Art History*, no. 3 (July). https://doi.org/10.11588/dah.2018.3.49915.

7. Brey, Alexander. 2021. "Digital Art History in 2021." *History Compass* 19 (8): e12678. https://doi.org/10.1111/hic3.12678
8. Caddy, Scott Allen. 2021. "The Significance of Literary Outliers in Nineteenth-Century British Fiction: A Stylometric Analysis." PhD diss., Arizona State University. https://hdl.handle.net/2286/R.2.N.161483
9. Chiquet, Vera, Lucas Burkart, Peter Fornaro, and Jane Haller, eds. 2024. *WORKFLOWS: Digitization Projects in GLAM and Research Institutions*. Zenodo. https://doi.org/10.5281/zenodo.11501686.
10. Chu, Kah Em, Pantea Keikhosrokiani, and Moussa Pourya Asl. 2022. "A Topic Modeling and Sentiment Analysis Model for Detection and Visualization of Themes in Literary Texts." *Pertanika Journal of Science & Technology* 30 (4): 2535–2561. https://doi.org/10.47836/pjst.30.4.14
11. Cuppen, Iris, and Inge van de Ven. 2019. "DHQ: Digital Humanities Quarterly." *DHQ* 13 (3). https://www.digitalhumanities.org/dhq/
12. DeepSeek. 2025. *Models & Pricing | DeepSeek API Docs*. Accessed April 16, 2025. https://api-docs.deepseek.com/quick_start/pricing.
13. Dilai, Marianna, and Iryna Dilai. 2024. "Literary Text Mining Using Verb Feature Clustering." https://ceur-ws.org/Vol-3723/paper15.pdf
14. Do, Thao, Dinh Phu Tran, An Vo, and Daeyoung Kim. 2025. "Reference-Based Post-OCR Processing with LLM for Precise Diacritic Text in Historical Document Recognition." *AAAI Conference on Artificial Intelligence*. https://github.com/thaodod/VieBookRead.
15. Douglas, Will. "The Inside Story of How ChatGPT Was Built—from the People Who Made It." *MIT Technology Review*, March 3, 2023. https://www.technologyreview.com/2023/03/03/1069311/inside-story-oral-history-how-chatgpt-built-openai/.
16. Drucker, Johanna. 2013. "Is There a 'Digital' Art History?" *Visual Resources* 29 (1–2): 5–13. https://doi.org/10.1080/01973762.2013.761106.
17. Drucker, Johanna, Anne Helmreich, Matthew Lincoln, and Francesca Rose. 2015. "Digital Art History: The American Scene." *Perspective* 2. https://doi.org/10.4000/perspective.6021.
18. Fisher, Michelle Millar, and Anne Swartz. 2014. "Why Digital Art History?" *Visual Resources* 30 (2): 125–137. https://doi.org/10.1080/01973762.2014.925410.
19. Fleischhacker, David, Wolfgang Goederle, and Roman Kern. 2024. "Improving OCR Quality in 19th Century Historical Documents Using a Combined Machine Learning-Based Approach." *arXiv* preprint arXiv:2401.07787. https://arxiv.org/abs/2401.07787.
20. Gabay, Simon, Barbara Topalov, Caroline Corbières, Lucie Rondeau Du Noyer, Béatrice Joyeux-Prunel, and Laurent Romary. 2021. "Automating Artl@s – Extracting Data from Exhibition Catalogues." In *EADH 2021: Second International Conference of the European Association for Digital Humanities*. https://hal.science/hal-03331838v1
21. Hoetjes, Lotte. 2025. "Training Transkribus on Middle Dutch: Are AI Transcriptions of Handwritten Text Here to Stay?" *Hasta – St Andrews Institute for Intellectual*

History, April 7, 2025. https://www.hasta-standrews.com/features/2025/4/7/training-transskribus-on-middle-dutch-are-ai-transcriptions-of-handwritten-text-here-to-stay.

22. Hyvönen, Esko. 2020. "Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data Analysis and Serendipitous Knowledge Discovery." *Semantic Web* 11 (1): 187–193. https://doi.org/10.3233/SW-190386.
23. Impett, Leonardo, and Fabian Offert. 2022. "There Is a Digital Art History." *Visual Resources* 38 (2): 186–209. https://doi.org/10.1080/01973762.2024.2362466.
24. Jaskot, Paul B. 2019. "Digital Art History as the Social History of Art: Towards the Disciplinary Relevance of Digital Methods." *Visual Resources* 35 (1–2): 21–33. https://doi.org/10.1080/01973762.2019.1586951.
25. Khan, Arsh, Utsav Rai, Shashank Shekhar Singh, Yukinori Yamamoto, Xabier Granja Ibarreche, Harrison Meadows, and Sergei Gleyzer. 2024. "OCR Approaches for Humanities: Applications of Artificial Intelligence/Machine Learning on Transcription and Transliteration of Historical Documents." *Digital Studies in Language and Literature* 1 (1–2): 85–112. https://hal.science/hal-03331838v1
26. Khemakhem, Mohamed, Simon Gabay, Béatrice Joyeux-Prunel, Laurent Romary, Léa Saint-Raymond, and Lucie Rondeau Du Noyer. 2020. "Information Extraction Workflow for Digitised Entry-Based Documents." In *DARIAH Annual Event 2020*. https://hal.science/hal-02508549v1
27. Klarenbeek, Hanna. 2012. *Penseelprinsessen & Broodschilderessen: Vrouwen in de Beeldende Kunst, 1808–1913*. Bussum: Thoth.
28. Koot, Roman, and Jeroen Kapelle. 2014. "Catalogi Levende Meesters Gedigitaliseerd / Living Masters Catalogues Digitised." RKD – Netherlands Institute for Art History. https://rkddb.rkd.nl/rkddb/digital_book/202003113.pdf.
29. Lian, Yuntao, and Jiafeng Xie. 2024. "The Evolution of Digital Cultural Heritage Research: Identifying Key Trends, Hotspots, and Challenges through Bibliometric Analysis." *Sustainability* 16 (16): 7125. https://doi.org/10.3390/su16167125.
30. Luo, Yuxuan, Jiaqi Tang, Chenyi Huang, Feiyang Hao, and Zhouhui Lian. 2025. "CalliReader: Contextualizing Chinese Calligraphy via an Embedding-Aligned Vision-Language Model." *arXiv* preprint arXiv:2503.06472. https://arxiv.org/abs/2503.06472.
31. Malladhi, Avinash. 2023. *"*Transforming Information Extraction: AI and Machine Learning in Optical Character Recognition Systems and Applications Across Industries.*" International Journal of Computer Trends and Technology* 71, no. 4 (2023): 81–90.
32. Manovich, Lev. 2015. "Data Science and Digital Art History." *International Journal for Digital Art History*, no. 1. https://doi.org/10.11588/dah.2015.1.21624.
33. Maxwell, Michael, and Aric Bills. 2017. "Endangered Data for Endangered Languages: Digitizing Print Dictionaries." In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 118–127. https://aclanthology.org/W17-0113/
34. Mechi, Olfa, Maroua Mehri, Rolf Ingold, and Najoua Essoukri Ben Amara. 2020. "A Text Line Extraction Method for Archival Document Transcription." In *2020 17th*

International Multi-Conference on Systems, Signals & Devices (SSD), 479–484. IEEE. https://doi.org/10.1109/SSD49366.2020.9364203.

35. OpenAI. 2023. "Introducing ChatGPT." *OpenAI*. Accessed April 17, 2025. https://openai.com/blog/chatgpt.
36. OpenAI. 2025a. Model - OpenAI API. Accessed April 16, 2025. https://platform.openai.com/docs/models/gpt-4o.
37. OpenAI. 2025b. Model - OpenAI API. Accessed April 16, 2025. https://platform.openai.com/docs/models/gpt-4-turbo.
38. OpenAI. 2025c. *Pricing*. Accessed April 16, 2025. https://openai.com/api/pricing/.
39. Poulopoulos, Vassilis, and Manolis Wallace. 2022. "Digital Technologies and the Role of Data in Cultural Heritage: The Past, the Present, and the Future." *Big Data and Cognitive Computing* 6 (3): 73. https://doi.org/10.3390/bdcc6030073.
40. RKD – Netherlands Institute for Art History. 2025. "About the RKD." Accessed April 16, 2025. https://www.rkd.nl/en/about-the-rkd.
41. RKD – Netherlands Institute for Art History. 2025. *RKD Research*. Accessed April 16, 2025. https://research.rkd.nl/en.
42. Rodríguez Ortega, Nuria. 2013. "Digital Art History: An Examination of Conscience." *Visual Resources* 29 (1–2): 129–133. https://doi.org/10.1080/01973762.2013.761116.
43. Ros, Rasmus, Elizabeth Bjarnason, and Per Runeson. 2017. "A Machine Learning Approach for Semi-Automated Search and Selection in Literature Studies." In *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering*, 118–127. Karlskrona, Sweden: ACM. https://doi.org/10.1145/3084226.3084243.
44. Salgaro, Massimo. 2023. *Stylistics, Stylometry and Sentiment Analysis in German Studies: The Operationalization of Literary Values*. 1st ed. Göttingen: V&R Unipress. https://doi.org/10.14220/9783737015707.
45. Scheithauer, Hugo, Sarah Bénière, and Laurent Romary. 2024. "Automatic Retro-Structuration of Auction Sales Catalogs Layout and Content." In *DH2024: Reinvention and Responsibility*. Washington, D.C.: Alliance of Digital Humanities Organizations. https://hal.science/hal-04547239.
46. Slocombe, Will, and Genevieve Liveley, eds. 2024. *The Routledge Handbook of AI and Literature*. 1st ed. New York: Routledge. https://doi.org/10.4324/9781003255789.
47. Sobchuk, Oleg, and Artjoms Šeļa. 2024. "Computational Thematics: Comparing Algorithms for Clustering the Genres of Literary Fiction." *Humanities and Social Sciences Communications* 11 (1): 1–12. https://doi.org/10.1057/s41599-023-02138-4.
48. Solkiran, Mert. "Beyond OCR: How LLMs Are Transforming Structured PDF Extraction." *LinkedIn*, February 9, 2025. https://www.linkedin.com/pulse/beyond-ocr-how-llms-transforming-structured-pdf-mert-solkiran-8c1nf/.
49. Smith, Virginia Carlson, and William R. Treese. 1975. "A Computerized Approach to Art Exhibition Catalogs." *Library Trends*. Urbana-Champaign: Graduate School of Library and Information Science, University of Illinois. http://hdl.handle.net/2142/6801.
50. Sotirova, Kalina, Juliana Peneva, Stanislav Ivanov, Rositza Doneva, and Milena Dobreva. 2012. "Digitization of Cultural Heritage – Standards, Institutions,

Initiatives." In *Access to Digital Cultural Heritage: Innovative Applications of Automated Metadata Generation*, 23–68.

51. Storm, Scott, and Emily C. Rainey. 2024. "Form, Criticality, and Humanity: Topic Modeling the Field of Literary Studies for English Education." *English Teaching: Practice & Critique* 23 (3): 388–403. https://doi.org/10.1108/ETPC-11-2023-0151.
52. Swiss German University. 2025. "A Comparison of Leading AI Models: DeepSeek AI, ChatGPT, Gemini, and Perplexity AI." *Swiss German University*. Accessed April 17, 2025. https://sgu.ac.id/a-comparison-of-leading-ai-models-deepseek-ai-chatgpt-gemini-and-perplexity-ai/.
53. Tibbe, Lieske. 2021. "Book Review of *Mirror of Reality: Nineteenth-Century Painting in the Netherlands* and *Spiegel van de werkelijkheid. 19de-eeuwse schilderkunst in Nederland* by Jenny Reynaerts." *Nineteenth-Century Art Worldwide* 20 (1). https://doi.org/10.29411/ncaw.2021.20.1.13.
54. Thomas, Alan, Robert Gaizauskas, and Haiping Lu. 2024. "Leveraging LLMs for Post-OCR Correction of Historical Newspapers." In *LT4HALA 2024 @ LREC-COLING*. https://aclanthology.org/2024.lt4hala-1.14/
55. Vangeli, Marius. 2024. "Large Language Models as Advanced Data Preprocessors: Transforming Unstructured Text into Fine-Tuning Datasets". Uppsala University. https://uu.diva-portal.org/smash/get/diva2:1879125/FULLTEXT01.pdf
56. Veninga, Martijn. 2024. "LLMs for OCR Post-Correction." MSc thesis, University of Twente. https://purl.utwente.nl/essays/102117
57. Weber, Josef Thilo. 2021. "Extracting Retrievable Information from Archival Documents". PhD diss., Technische Universität Wien. https://doi.org/10.34726/hss.2021.93623
58. Yadav, Deny. 2024. "The Role of Artificial Intelligence in Literary Analysis: A Computational Approach to Understand Literary Styles." *International Journal of Emerging Knowledge Studies* 3: 558–565. https://doi.org/10.70333/ijeks-03-09-006.
59. Zweig, Benjamin. 2015. "Forgotten Genealogies: Brief Reflections on the History of Digital Art History." *International Journal for Digital Art History*, no. 1 (June). https://doi.org/10.11588/dah.2015.1.21633.