Temporal Bias in Historical Newspaper Digitisation: Evidence from Reconstructing the British Mandate Trademark Registry

By

Eran Toch,[*] Bar Ifrah,[†] Michael Birnhack[‡]

Forthcoming: Journal of Documentation (2026)

Abstract

Purpose – Digitising historical resources has become a critical step in historical research, augmenting manual processes and facilitating big data studies. However, older material is of lower quality and may affect the quality of the digitisation. We document and quantify this bias in the context of a trademark registry reconstruction project, which involves identifying and matching graphical trademarks from notifications published in the Official Gazette during the British Mandate in Palestine (1917-1948).

Methodology – We develop a rule-based pipeline combining graphical object extraction and Optical Character Recognition (OCR) to match trademark images to metadata records across more than 300 Gazette editions, spanning 26 years. To measure the temporal dimension of digitisation quality, logistic regression is applied to the full dataset of 7,263 trademarks to isolate the effect of publication date on matching probability, independently of structural layout changes.

Findings – The pipeline achieves an overall identification rate of 86.6%. Using logistic regression, we find a statistically significant temporal effect: each additional year is associated with a 0.38 percentage-point increase in matching probability, corresponding to approximately 13% improvement over the 26-year study period. The decline in accuracy for 1920s editions is attributable to physical paper deterioration and scan quality rather than structural layout changes.

Research Limitations/Implications – The pipeline is tailored to the Palestine Gazette's specific conventions and relies on commercial OCR software, which limits direct replicability without adaptation. The study draws on a single jurisdiction and source type, so the magnitude of chronological bias may differ across other corpora.

[**] Professor, School of Industrial & Intelligent Systems Engineering, Tel Aviv University.
[††] Masters Student, School of Industrial & Intelligent Systems Engineering, Tel Aviv University.
[‡] Professor of Law, Faculty of Law, Tel Aviv University.
We wish to thank Raz Ashkenazi for leading the legal research team. Birnhack acknowledges the support of ISF Grant 532/21

Practical Implications – The rule-based extraction and matching pipeline provides a replicable, low-threshold tool for analogous digitisation projects, including other colonial trademark or patent registries and official gazettes, without requiring advanced machine-learning infrastructure. The reconstructed British Mandate trademark registry, now openly available, enables historical and comparative research on commercial activity, legal history, and colonial governance. The finding of a chronological bias urges digitisation project managers to incorporate temporal bias measurement into quality-assurance workflows and to stratify or weight findings by publication period rather than aggregating results uniformly across time.

Originality/value – Our research reveals that the deteriorating efficiency of a newspaper-analysing algorithm over time may lead to a chronological bias, skewing historical data toward more recent periods, a critical factor we aim to document and highlight to support more accurate and balanced historical research. The research tool we developed facilitates similar digitisation projects for trademarks and other images, along with their accompanying information. This is a necessary step in reconstructing old or lost trademark registries, which can then contribute to understanding the economic and cultural dimensions of the relevant jurisdiction.

Introduction

Digital texts have become essential tools for humanities and historical research, making the quality assessment of scanned resources crucial for understanding the quality and validity of digital humanities infrastructure and collections. The digitisation of printed historical materials, such as newspapers, preserves historical information, provides digital access to these collections, and enables the building of corpora for digital humanities research (Bremer-Laamanen, 2006; Rydberg-Cox, 2009; Gooding, 2016; Neudecker et al., 2019). The digitisation of historical documents, which contain both textual and visual elements, presents unique challenges beyond traditional text recognition. Large-scale newspaper digitisation projects, such as Europeana (Pletschacher et al., 2015) and the Aftonbladet digitisation (Jarlbrink & Snickars, 2017), primarily focused on improving Optical Character Recognition (OCR) accuracy and correcting post-processing errors to create searchable text archives. Similarly, the Finnish newspaper digitisation project (Drobac et al., 2017) emphasised improving OCR performance, with success measured by how closely digital replicas matched original print content. Other uses of automated and machine-learning solutions in creating and managing digital archives include handwriting recognition (Nockels et al., 2024) and metadata extraction (Aske & Giardinetti, 2023; Arnold et al., 2017).

In this paper, we discuss newspaper digitisation in the context of the reconstruction of the British Mandate's trademark registry, from the Official Gazette (renamed the Palestine Gazette in 1932) published by the local British government during the British Mandate in Palestine from 1917 to 1948 (Birnhack, 2023). Our objective was to create and evaluate a reliable method for automatically extracting trademark images from scanned newspaper pages and matching them with existing metadata records in a structured database. This study examines the quality of digitisation for a specific historical collection. During this era, the British Government frequently published notifications of trademark registration applications in the Gazette, either sporadically or

within dedicated sections, resulting in over 300 such publications between 1922 and 1948. Extracting and matching graphical elements enables comprehensive searchability of visual trademark data that would otherwise remain buried in scanned documents, allowing researchers to trace the evolution of commercial branding and business practices during the Mandate period.

The digitisation of newspaper multimedia content is inherently complex because it involves not only text recognition but also image analysis, layout interpretation, segmentation, metadata generation, and quality control, all of which require a combination of automated systems and human oversight. Several case studies have pointed to challenges in managing printed text and visual elements, such as identifying front-page covers in newspapers in the National Library of Norway's collections (de la Rosa, 2025), article-level segmentation in KBR's BelgicaPress collection (Ali et al., 2024), or extraction and visual grouping of printed illustrations in digitised chapbooks in Scotland (Dutta et al., 2021).

While OCR tools can distinguish between text regions and image areas in newspaper layouts, extracting graphics is only the first step—the more complex challenge lies in computationally associating each extracted image with its corresponding textual metadata, particularly when layout conventions change over time. These complications are compounded by the inherent characteristics of historical newspapers, which have long challenged OCR systems: variable typography, multiple-column layouts, and complex page structures that create what Conway (2013) termed "remediation layers," where scanning, OCR, and format conversion each introduce potential errors. The problem intensifies when the relationship between visual and textual elements is neither fixed nor explicit—as in trademark notices where the image may appear above, below, or even on a different page from its descriptive text.

One of the main questions that arises from automated extraction and matching processes is that the condition of historical materials may significantly influence digitisation quality, creating particular challenges for older documents, which are central to humanities research. For example, Tanner et al. (2009) demonstrated this relationship empirically by measuring character accuracy, word accuracy, and significant word accuracy using the British Library's 19th-century newspapers database, revealing how temporal factors affect OCR performance. Similarly, Strange et al. (2014) found that OCR errors correlate with different types of newspaper content, though their study did not control for publication date. Hill and Hengchen (2019) used the large-scale Eighteenth Century Collections Online (ECCO) for similar purposes. The effects of scanning and analysis procedures are not merely technical: As Cordell (2017) argued, mass-digitised texts should be understood as distinct bibliographic objects, whose OCR layers introduce structured distortions that shape what can be found and, therefore, what can be studied.

These findings suggest that the physical deterioration of older materials, combined with historical printing technologies and the quality of paper, has compounding effects on digitisation accuracy. Our dataset enables better control over temporal variables, as the scanning quality and the newspaper's structure remain consistent throughout our analysis. Moreover, rather than questioning the quality of the overall text scanning, which relies on spelling corrections and other methods, as Van Strien et al. demonstrated (2020), our method is relevant to metadata matching with extracted graphical elements. As Jarlbrink and Snickars (2017) have warned, such quality issues can harm the documentation process itself, potentially introducing systematic biases that affect scholarly interpretation and research outcomes based on these digital collections.

Background

Trademark Data

Studying the history of trademark law provides a roadmap of legal evolution, offering invaluable insights into how business norms, regulations, and standards have evolved and changed over time (Bently & Bone, 2024). The overall trademark registration in a particular jurisdiction can often serve as an economic indicator, shedding light on market trends, competition in particular industries, and business activity within certain periods (Duguid et al., 2010). Beyond their commercial relevance, trademarks reflect sociocultural values and the political context, with the potential to provide a deeper understanding of how societal norms evolve and transform (see, e.g., Scardamaglia, 2015, on trademarks in colonial Australia). The trademarks are subject to a semiotic analysis, which is used to portray linguistic aspects of commodities (Beebe, 2003). Ultimately, historical preservation enriches our understanding of the past, offering key insights across fields such as law, economics, history, and sociology.

Modern trademarks emerged in the late 19th century, alongside the Second Industrial Revolution (Sherman & Bently, 1999: 166-172; Bently, 2008). Trademarks are signs attached to goods that indicate their origin and serve as intermediaries between manufacturers or traders and end consumers. Along with patents, trademarks play a crucial role in the market, communicating important information, reducing consumers' search costs (Landes & Posner, 1988), incentivising incumbent market players to maintain consistent product quality, and building reputation. In this regard, trademarks are a market tool and a form of intellectual property.

Accordingly, non-physical trademarks can hold significant value for a business (Corbett et al., 2008). Signs such as logos, phrases, and colours, like McDonald's golden arches or the Coca-Cola emblem, encapsulate the goodwill and reputation of manufacturers and service providers. Trademarks convey information about their origin, the quality of the product or service (which may be high or low; consistency is what matters here), and often instil perceptions of quality, safety, and in some cases, a sense of community among consumers (Smith & Richey, 2013). Some trademarks are not only signifiers but also have commercial value, thereby becoming brands (Wilkins 1992) and often being framed as property, following suggestions made a century ago (Schechter 1925).

Given the economic benefits of trademarks, the law provides protection against unauthorised use of the marks through a government examination and registration system. Legal protection is needed so that the trademark owner can prevent a competitor from imitating the trademark in a manner that deceives consumers. If a consumer relies on the trademark but is deceived, the imitation both misleads the consumer and undermines the entire trademark system. Trademark law governs this field, determining which trademarks are eligible for legal protection, what constitutes infringement, and the appropriate remedies.

To enjoy legal protection under trademark law, manufacturers or traders must register their marks with the governmental registry, typically operated by the Patent and Trademark Office (PTO). The first registries were founded in the 19th century in France (Duguid, 2009) and some of the self-governing dominions of the British Empire, namely Australian colonies (Bently, 2011: 170). In 1875, the British established their own trademark registry (Bently, 2011).

The British took over Palestine in 1917-18, and ended 400 years of Ottoman rule. In the summer of 1920, the British established a civil government in Jerusalem, which began reviewing the local legislation (Likhovski, 2006). Trademark law was on the legislative checklist, and in late 1921, the British enacted a local Trademarks Ordinance, which took effect in January 1922 (Birnhack, 2021). A trademark registration was established based on the 1875 British model and the 1907 British Trademark Act. The local British Official Gazette frequently published trademark registration notifications, either sporadically or within a specific segment. Throughout the Mandate period, over 300 such publications were published (1922-1948).

After the fees were paid, the PTO examined the trademark applications. If it was satisfied that the application met the legal standard and that there was no reason to refuse registration, it was published in the Official Gazette (renamed the Palestine Gazette in 1932), allowing a short window for pre-grant oppositions. If no opposition were submitted, or if opposition was submitted but dismissed, the application was sealed, and the trademark enjoyed legal force. The original trademark registration was lost, necessitating its reconstruction; the reconstruction was based on the publications of the trademark applications (Birnhack, 2023). Before attending to our technological development, we explain the importance of trademark data for historical research.

Gathering trademark data can provide various indications of the trademark activity, opening up numerous research questions. First, referring to the metadata, we can observe the overall number of applications over the years or in a particular industry, based on the statutory classification of the marks. We can examine the country of origin of the applicants and further review a country's submissions over time or in a particular industry. Such inquiries raise questions such as the correlation between trademark submissions and legislative changes, economic changes, and geopolitical events. The metadata enables us to compare trademark activity in a specific industry with the overall pattern or with other sectors, suggesting interesting insights for business historians. Comparing applications across countries indicates which countries used the system and which did not. The intensity of trademark activity in a particular industry may indicate the level of competition therein, and focusing on specific applicants can teach us about that business or that individual.

Trademark data also contains the marks themselves. We can inquire about the ratio of trade words versus images; for the trade words, we can ask about the languages used, and for the images, we can undertake a discourse analysis. There was a flow of trademarks into the country via imports and out of the country, for exports. Trademarks may reflect social, cultural, religious, or national ideologies, and provide a fresh resource for the imagery of the place and time under inspection.

The trademark registry illustrates how historical archives can highlight commercial and cultural practices in a colonial setting (Candela et al., 2023). Whereas archives typically contain the coloniser's side, the historical trademark data reveals on-the-ground practices, adding the colonised's reception of the foreign-imposed tool. Established by the coloniser to serve first and foremost its own interests, the trademark registry may be a successful legal and administrative transplant, but it might also meet local indifference, selective usage, or even resistance. Analysing these records thus provides a more grounded account of how colonial law was received, negotiated, and operationalised, complicating assumptions about its reach and effectiveness.

In Mandate Palestine's trademark registry, reconstructing it based on published applications raised numerous challenges, including changes in the numbering system, discontinuous applications, and transitions from the Ottoman period to the British, and then from the British to the Israeli regime. Here we focus on the technical – but crucial – issue of extracting the marks from the black and white, old and not always clearly printed, official gazettes, and associating the images with the meta-data about the applicant, their country of origin, date of application, the number given to the application, and occasionally some other additional information that was contained in the published application. The reconstructed registration is open and available to all, now alongside a reconstructed patent registry of the same period (Birnhack, 2023).

Extracting Structured Metadata from Historical Document Collections

Extracting structured metadata from digitised historical newspapers poses challenges beyond optical character recognition. Even when OCR quality is high, the spatial relationships between a document's textual and visual elements remain implicit in the page image; recovering them requires tools capable of analysing document structure, identifying and classifying page regions, and linking extracted content across a collection (Philips and Nasseh, 2020). Recent scholarship has shown that OCR-focused pipelines require combining text processing with image segmentation, as OCR output alone does not resolve the spatial and structural relationships between articles and illustrations on a page. The Impresso project (Ehrmann et al., 2020) exemplifies this challenge, as it required manual annotation and image location during the digitisation of 200 years of multilingual newspaper material from Switzerland and Luxembourg.

Ali et al. (2024), working with KBR's BelgicaPress collection, demonstrate a four-stage enrichment workflow: article-level segmentation using ABBYY FineReader output, extraction of specific article types, image clustering, and named-entity recognition linked to open data. Their finding that OCR quality directly constrains the performance of every downstream stage confirms the bottleneck described by Jarlbrink and Snickars (2017) and Hill and Hengchen (2019). Their workflow addresses a closely related but distinct problem from ours: clustering images into categories to improve collection searchability, whereas we perform one-to-one matching of extracted trademark images to pre-existing records in an external database.

Approaches to structured information extraction from historical document collections can be broadly divided into rule-based methods, which apply explicit patterns and spatial heuristics to OCR output (Chiticariu et al., 2013), and machine learning methods, which learn to identify and classify document elements from annotated training data, as demonstrated by transformer-based document understanding models (Xu et al., 2020) and ML-driven content extraction pipelines such as those of Dutta et al. (2021) and Ali et al. (2024). The choice between them depends on whether the structural regularities of the source material can be explicitly described and on whether sufficient labelled training data exists for the domain in question. The challenge is compounded in our case by the variable spatial relationship between trademark images and their associated notices, which shifted across the Gazette's publication history.

A related strand of tooling has developed within Automated Text Recognition (ATR) platforms more broadly. Tools such as Transkribus (Kahle et al., 2017) and eScriptorium (Kiessling et al., 2019), widely used for historical handwriting and print recognition in the humanities, incorporate structural segmentation models that classify page regions, distinguishing text blocks from image

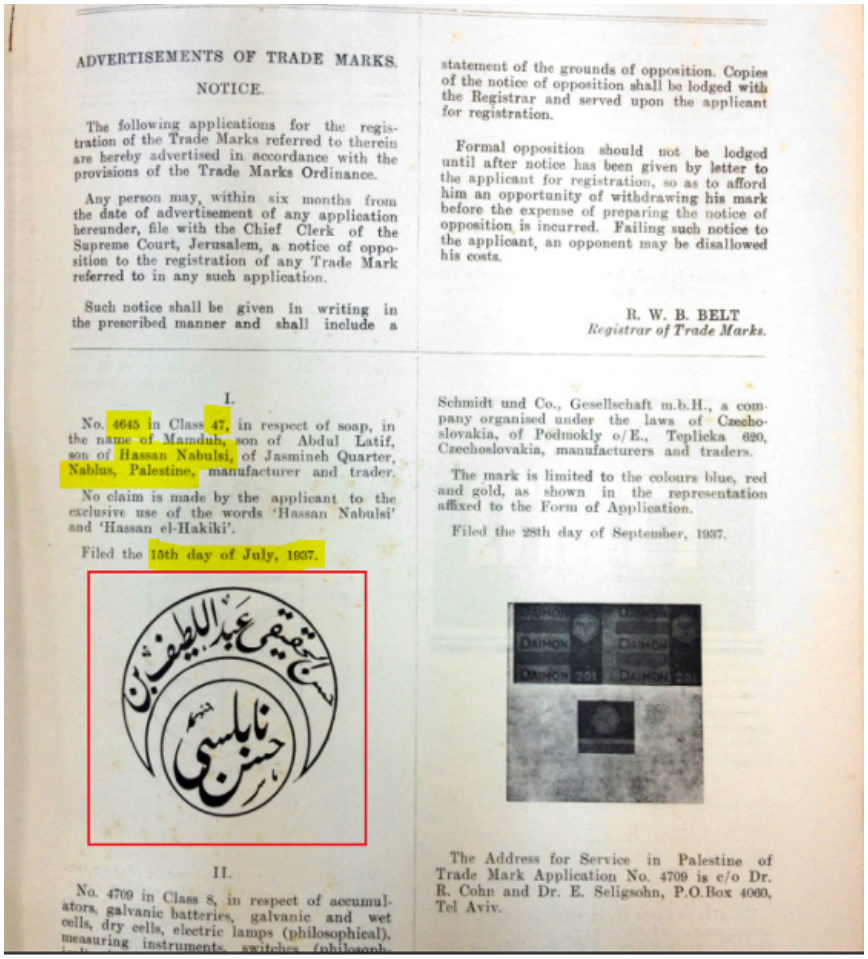
areas, and export results in XML-based formats. This structural layer provides a framework for marking up and preserving image regions alongside OCR text, though image recognition in these platforms operates at the level of layout rather than the semantic content of the visual elements themselves. However, as our method demonstrates, structural segmentation is only a precondition for the more demanding task of algorithmically associating each extracted image with its corresponding textual metadata across a document collection spanning 26 years and multiple layout conventions.

Newspaper digitisation workflows are not purely automated pipelines but hybrid systems in which manual labour continually feeds back into and improves computational processes. Human correction of OCR errors, validation of segmentation, and review of image–text matching directly shape training data and extraction rules. Yet successful digitisation requires more than technical accuracy: it requires interfaces that make collections meaningfully explorable. Search-driven systems support precise queries but often fail to accommodate the exploratory, pattern-seeking practices central to humanities research (Ruecker et al., 2016). Scholars have therefore called for interfaces that foreground collection structure, scale, and interpretive engagement rather than hiding complexity behind a single search box (Drucker, 2020). These interfaces enable easy browsing and creation of connections between data and its context (Whitelaw, 2015), as well as the construction of workflows in which data flows through different tools without losing context (Turkel et al., 2012). For reconstructed trademark registries, where patterns across time, industries, and origins are not known in advance, the central challenge is not only to build reliable matching algorithms but also to design environments in which researchers can navigate and interpret thousands of metadata records spanning decades.

Extracting and matching trademark data

Document preprocessing and OCR

Digitising the trademark notifications along with their associated trademark images can be quite complex. The primary challenges include handling poor-quality and distorted scans, as well as adapting to the Gazette's evolving structure over time. Throughout this article, we distinguish three levels at which digitisation challenges arise in documents like Mandate Palestine's Gazette. The structural level concerns the spatial organisation of the document: column arrangement, the positioning of trademark images relative to their corresponding notices, and layout conventions that changed over time (Cohen et al., 2016). The form level concerns the physical and typographic qualities of the material: paper condition, printing technology, scan resolution, and the cumulative effects of physical deterioration on legibility. The content level concerns the informational layer: OCR recognition of keywords, numbers, proper names, and application dates. As we show, each level of challenge requires distinct methodological responses, and the interaction between form degradation and content recognition is the primary driver of the chronological bias we report.



ADVERTISEMENTS OF TRADE MARKS.

NOTICE.

The following applications for the registration of the Trade Marks referred to therein are hereby advertised in accordance with the provisions of the Trade Marks Ordinance.

Any person may, within six months from the date of advertisement of any application hereunder, file with the Chief Clerk of the Supreme Court, Jerusalem, a notice of opposition to the registration of any Trade Mark referred to in any such application.

Such notice shall be given in writing in the prescribed manner and shall include a statement of the grounds of opposition. Copies of the notice of opposition shall be lodged with the Registrar and served upon the applicant for registration.

Formal opposition should not be lodged until after notice has been given by letter to the applicant for registration, so as to afford him an opportunity of withdrawing his mark before the expense of preparing the notice of opposition is incurred. Failing such notice to the applicant, an opponent may be disallowed his costs.

R. W. B. BELT
Registrar of Trade Marks.

I.

No. 4645 in Class 47, in respect of soap, in the name of Mamduh, son of Abdul Latif, son of Hassan Nabulsi, of Jasmineh Quarter, Nablus, Palestine, manufacturer and trader.

No claim is made by the applicant to the exclusive use of the words 'Hassan Nabulsi' and 'Hassan el-Hakiki'.

Filed the 15th day of July, 1937.

II.

No. 4709 in Class 8, in respect of accumulators, galvanic batteries, galvanic and wet cells, dry cells, electric lamps (philosophical), measuring instruments, switches, (philosoph-

Schmidt und Co., Gesellschaft m.b.H., a company organised under the laws of Czechoslovakia, of Podmokly o/E., Teplicka 620, Czechoslovakia, manufacturers and traders.

The mark is limited to the colours blue, red and gold, as shown in the representation affixed to the Form of Application.

Filed the 28th day of September, 1937.

The Address for Service in Palestine of Trade Mark Application No. 4709 is c/o Dr. R. Cohn and Dr. E. Seligsohn, P.O.Box 4060, Tel Aviv.

Figure 1: An example of a trademark application on the Mandate Palestine Gazette No. 802 of 28 July 1938, p. 79. The red frame indicates the trademark, and the highlighted text indicates related data.

Our goal was to automatically extract trademark images from the Gazette and match them with an application record in an existing database (manually extracted from the Gazette), with a focus on minimising the error rate in classification§. For instance, in Figure 1, one application is displayed as part of a newspaper page that contains other, unrelated matters. The pertinent data that was manually extracted, highlighted in yellow, included the application number, class number, applicant, application date, country, city, and the image itself. In this context, an accurate match involves associating the application data with the trademark image. The data in our existing dataset included:

- **Item No.**
- **Date of Application:** 15.7.1937 (July 15, 1937)
- **Applicant:** Mamduh, son of Abdel Latif, son of Nabulsi
- **City:** Nablus
- **Country of applicant**: Local

§ The code for the project is publicly available at: <anonymized>

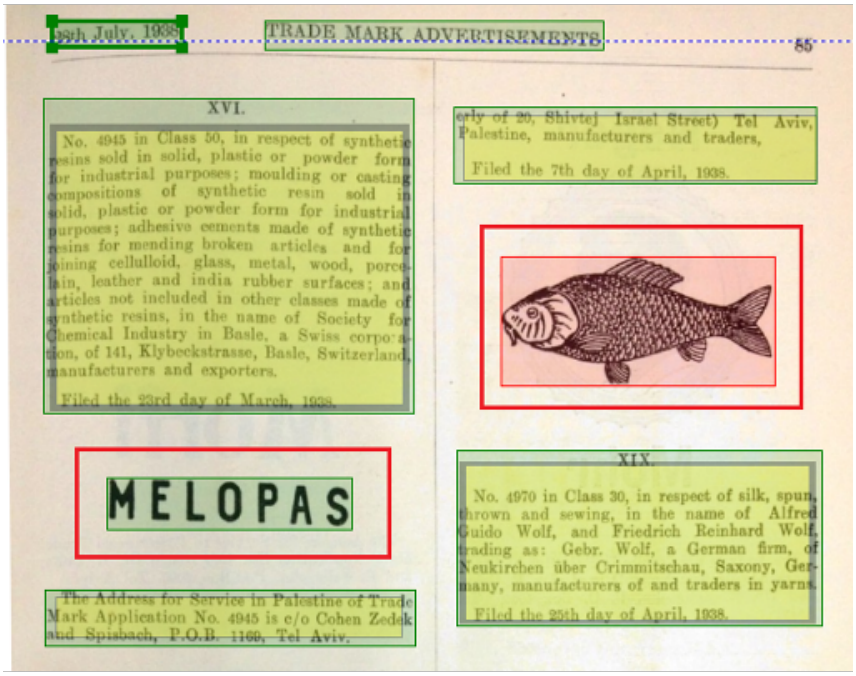


Figure 2: An example of object identification on the Mandate Palestine Gazette No. 802 of 28 July 1938, p. 85. The red frames stand for objects detected as images, and the green frames stand for objects detected as text areas. The red rectangles mean that this area should be recognised as an image. The yellow mark means that this area should be recognised as text.

The process for each paper requires six steps:

1. Pre-process the scanned Gazette's PDF file to generate a structured, readable file containing both text and images.
2. Extract all trademark images.
3. Gather as much data and metadata as possible.
4. Establish a match between each image and its corresponding data.
5. Compute the accuracy rate (if feasible) and identification rate.
6. Incorporate the results into the database.

Transforming an image-only PDF file into a searchable text PDF (where the text is interpreted as text and not an image) necessitated the use of OCR. The choice of OCR engine reflects broader trade-offs between commercial and open-source tools in cultural heritage digitisation. Comparative evaluations show that commercial systems such as ABBYY FineReader typically achieve higher out-of-the-box accuracy on degraded historical material and complex layouts, and have proved effective for processing old newspapers (Holley, 2009), while open-source engines such as Tesseract (Smith, 2017) can perform well on clean print but often require substantial training and post-correction to reach comparable reliability (Olson & Berry, 2021; Tafti et al., 2016). Our initial testing yielded similar results: open-source OCR tools identified only part of the label text. Moreover, the open-source tools we tested at the time of the study could not fulfil at least one core requirement: generating OCR files for papers of varying quality, handling old newspapers, and exporting to a structured format while retaining the paper's layout and image locations. Therefore, we chose ABBYY FineReader for preprocessing while implementing the

entity extraction and matching logic independently, resulting in a low-threshold workflow that can be adopted without advanced machine-learning infrastructure.

The results from ABBYY FineReader can be exported to various formats, such as HTML, DOCX, RTF, and other XML-based formats, and preserve the structure of the original paper. Its license is relatively inexpensive. When the tool is used to pre-process an image-only PDF, it automatically distinguishes between text and image areas. The text areas are highlighted with a green frame, and the image areas with a red frame, as depicted in Figure 2. For this task, we had to classify notices as text areas and trademark images as image areas. While the rest of the page data could be identified, it was not necessary for the subsequent steps.

Identifying a trademark

The rule-based design of our extraction pipeline was a deliberate methodological choice: the Palestine Gazette's layout follows a small set of explicit, stable geometric rules, column structure, vertical image–notice ordering, and a documented structural change in May 1928, that can be fully specified without training data, making a rule-based approach both appropriate and transparent (Chiticariu et al., 2013). The ground-truth dataset was reserved in its entirety for evaluation and could not be partitioned for ML training without compromising the assessment. The 86.6% identification rate achieved establishes a reproducible baseline against which future ML-based approaches can be benchmarked.

There are several methods to identify a paragraph as a notice in the output of the OCR software, an example of addressing content-level challenges. One approach could be to detect the Roman numeral at the start of the notice. However, due to OCR errors, the tool often failed to recognise the Roman numerals, rendering this approach ineffective. An alternative strategy involved identifying recurrent keywords. Words such as "No" (short for "number"), "class," and "In respect of" always appeared in the opening sentence of every notice. This latter approach proved more successful, with the main challenge being to address the spelling errors. Untreated post-OCR spelling errors can significantly degrade the quality of results. Therefore, we utilised prevalent pre-processing methods on the paragraphs, such as tokenisation, removal of dots and commas, and lowercase conversion (Uysal & Gunal, 2014).

Adopting a conservative approach that required all three keywords to be present in the sentence resulted in 50% of actual notices being misclassified. However, a more reasonable approach, which involved identifying at least two of the three keywords, increased the number of hits while maintaining a low number of false alarms in the confusion matrix. Despite the improvement, some cases were still misclassified. We, therefore, sought a method to manage spelling errors to increase the identification rate of the notices.

While post-OCR error-correction methods have been extensively researched, most published methods did not apply to our case. The goal was not to correct all errors in the paper, but to ensure specific keywords were recognised. Moreover, many of the spelling identification errors occurred in numbers, locations, or unique company names that lacked semantic information and could not be corrected by common language-based models. Consequently, we decided to use the edit distance method (Damerau, 1964). The edit distance is a metric that calculates the minimum number of operations needed to transform one string into another. This method allows us to calculate the edit distance between the keywords and the tokens in the paragraph. Initially, we utilised an external library to generate a list of all strings with an edit distance of 1 from each

keyword, then we examined whether the original keywords or any of the list items appeared in the paragraph. Eventually, we discovered that the best results were achieved with an edit distance of 1 for the keyword 'No', a frequently occurring term, an edit distance of 2 for "Class", and an edit distance of 2 for "In respect of". Utilising an edit distance of 3 did not alter the rate of identified notices but increased the number of false alarms, thereby reducing the accuracy rate.

Matching images and notices

Once we determined that a specific notice was indeed a trademark notice, the next phase involved extracting the relevant information from the notice. Extraction of the application number and the class number was relatively straightforward, given that the application number was always placed immediately after the word "No", and the class number followed the word "class". Upon identifying these fields, we verified whether the combination of application number and class number existed in the dataset rows for the specific publication date. Having compiled a list of all countries from the dataset, the algorithm determines if any of these countries were mentioned in the notice. We used a tactic to handle identification mismatches by rechecking with an edit distance of 2 for each country. The extraction of city and applicant information was accomplished using the same strategy. An external library facilitating the extraction of dates from text was used to derive the application dates.

The image-to-notice matching step addresses *structural-level* challenges and relies entirely on the Gazette's spatial layout rather than text recognition. Throughout the Mandate, particularly during the 1920s, numerous alterations were made to the Gazette's structure, which can be summarised into a set of guidelines. Until May 1928, the image was positioned above the notice (see Figure 3, left). Post-May 1928, the image was placed beneath the notice. Each page consisted of two equal columns, separated by a dividing line. The horizontal coordinate for each image/notice in the left column fell within the range $0 < x < 2{,}500$, whereas those in the right column exceeded 5000. When a notice was situated at the base of the left column, the related image would appear at the top of the right column (post-May 1928). When a notice appeared at the bottom of the right column, the associated image was found at the top of the left column on the subsequent page (post-May 1928). Exceptions (2-3%) were managed manually using ABBYY FineReader by dividing the original page into two sections.

Figure 3: Structure of Trademark graphics. Left: Trademark structure from the Palestine Gazette of 16 March 1928, where the image is shown above the notice. Right: an image (M.H.D.) from the Palestine Gazette of 28 July 1937 is positioned in the top right column of the next page when the notice is shown in the bottom right column of the previous page from the Palestine Gazette:

Information regarding the positions of the images and notices was compiled to serve as input for the matching algorithm. Since ABBYY FineReader conducted the image extraction process and it was manually validated, 100% of the images were successfully extracted. Each image carried data about its position within the document. We created a list of notices and information about their positions from the previous steps. As discussed above, each Gazette edition adheres to a set of rules concerning its structure. This algorithm applies this set of rules to pair a notice with its corresponding trademark image. The algorithm (post-May 1928) scans each notice, finds the list of candidate images in the same column that satisfy y_image > y_notice, and selects the image with the minimum y value. When this list is empty (when the notice is at the base of the column), it generates a list of all candidate images from the next column and selects the image with the minimum y value. Similar logic applies when the image appears on the subsequent page. This algorithm also carries out a validation step to prevent cumulative errors.

Algorithmic Optimisations

The algorithm we describe can undergo further optimisation. The extract function's first step is to generate a list of all unmatched rows. This list serves as the foundation for the filtering algorithm. This algorithm does not associate a notice with a row if it returns more than one potential candidate meeting all the filtering criteria.

If a second iteration of the extract function is applied, some notices from the first run that remained unpaired may find a match. For instance, consider a scenario where a company publishes five notices for the same product in a single newspaper. The only differing factor amongst them is the application number. During the initial run, the algorithm matches notices 1, 2, 4, and 5. Notice 3,

however, isn't matched, as the application number was undetected or partially detected, leading the algorithm to return multiple possible candidates.

When the extract function runs a second time, the remaining third notice gets paired with the leftover unmatched dataset row. This is because the other notices have already been matched, leaving no unpaired candidates. A similar scenario arises for low-quality scanned or blurry notices where the only available information is the country and the application's date. As the number of candidates decreases during the second or third run, the filtering algorithm may return only a single candidate row.

A critical hyperparameter employed was the verification level. At verification level 1, the procedure operates as previously explained. When the verification level is set to 2, the filtering function also applies the last filtering criterion: pattern matching. While this improves the identification rate, it reduces confidence in the matches. This pattern-matching criterion was found to be particularly beneficial for addressing the remaining blurry notices. By establishing two distinct verification levels, the identification rate can be maximised with a minimal impact on accuracy. The procedure commences at verification level 1, as previously described. After several runs, most notices are matched. The algorithm first targets the most informative notices, reserving the blurry ones for last. The last few notices that could not be identified due to low-quality scans are analysed using pattern matching. This method guesses the correct match. Since “guessing” is applied only to a small group of remaining notices, the chance of an incorrect match is significantly reduced.

Evaluation metrics

To evaluate the matching process, the first author manually examined the match and determined whether it was accurate. Similar to Ali et al. (2024), our two metrics for the classifier performance are the identification rate and accuracy rate. The identification rate (in percentage),reflects the number of matches between an image and a record relative to the total number of notices. The accuracy rate reflects the number of correct matches between an image and its notice, divided by the total matches made for that paper. For example, if a specific paper has 20 notices, 16 matches are made, and 15 of them are found to be correct in a manual examination, then the identification rate is 16/20 = 80%, and the accuracy rate is 15/16 = 93.75%. The matching algorithm’s objective was to automatically identify at least 80% of the trademarks while maintaining high accuracy (close to ~95%).

Findings

The initial dataset comprises 7,263 trademark rows from more than 300 editions published between 1922 and 1948 (Birnhack, 2021). Out of these, 7,208 rows have a known publication date. The identification rate results grouped by the Gazette’s publication decade are presented in Table 1 and Figure 4. The results showcase the identification rate (number of matches out of total notices published) and the accuracy rate (number of correct matches out of total matches), calculated for each paper in the sample. For the remaining documents, only the identification rate was determined. The database includes rows for notices of trademark applications from 1922 (when the trademark Ordinance took effect and official registration commenced) to 1948 (when the

British Mandate ended). The number of applications per year varied drastically from year to year, as can be seen in Figure 4. The reasons are detailed by Birnhack (2021), but this variance does not affect the OCR matching rate.

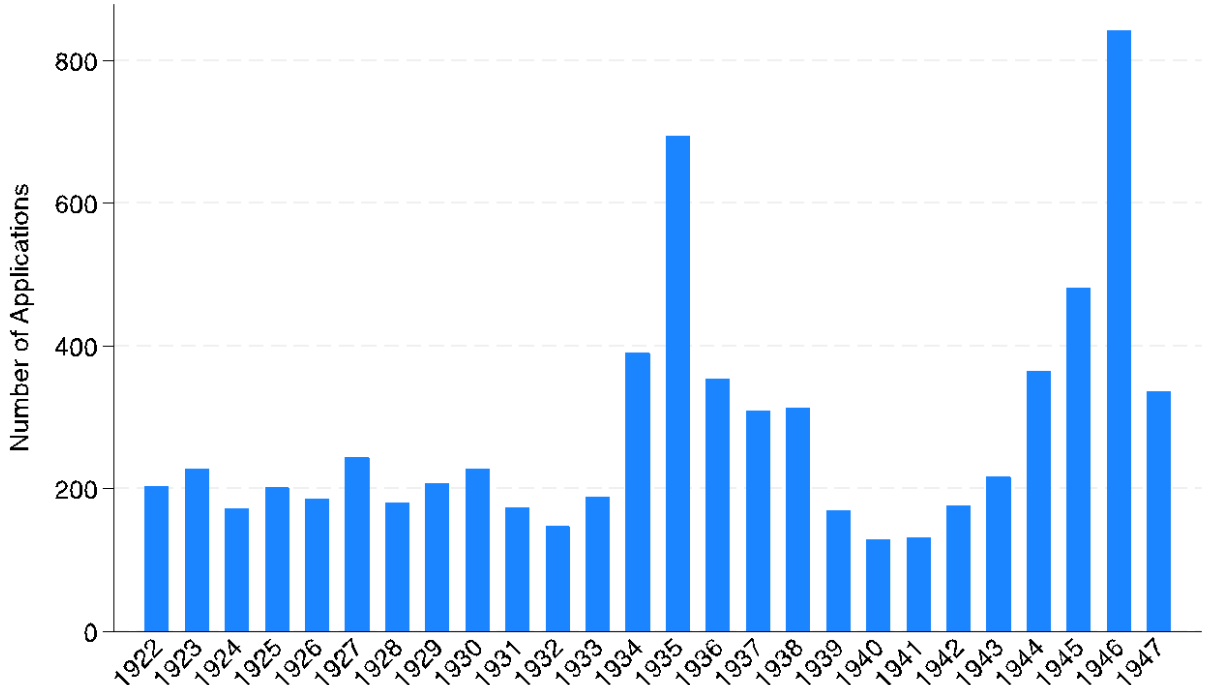


Figure 4: The number of trademark applications published per year.

The quality of the scan can limit the reliability of the filtering algorithm's matching criteria, resulting in a lower identification rate. This issue is the primary cause for the drop in identification rates for Gazette editions from the 1920s. If an edition was not manually analysed, its identification rate was computed without the accuracy rate, leading to consistently high identification rates across most papers.

The average identification rate was 86.6%, meaning 86.6% of the rows were successfully matched to the correct trademark image. Our analysis examines the factors influencing the likelihood of trademark applications appearing in newspapers, as identified through OCR. Overall, we observe a 13.34% increase in matching over our 26 years of sampling. Using logistic regression on 7,263

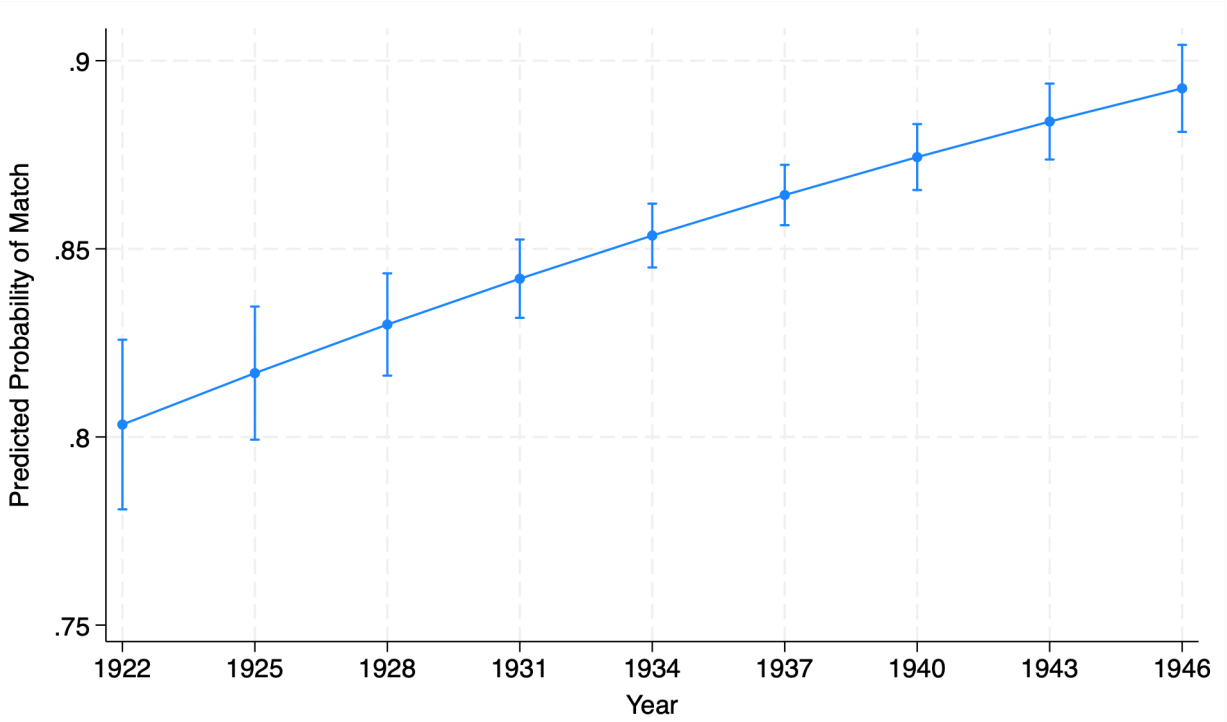


Figure 5: The marginal predicted probability of matching an application using OCR according to a binary logistic regression.

observations, we find a significant temporal marginal effect. As visualized in Figure 5, the probability of matching increased steadily over the study period. Each additional year was associated with a 0.38 percentage point increase in matching probability ($p < 0.001$), rising from approximately 80% in 1922 to 89% in 1946. The model is statistically significant (LR $\chi^2 = 43.14$, $p < 0.001$). Table 1 presents a breakdown of the average identification rate by decade, showing improvements in identification in each consecutive decade.

Decade	Number of trademarks	Matched trademarks	Average identification rate (%)	Average Accuracy rate (%)
1920s	1511	1279	84.65%	93.97%
1930s	2944	2555	86.79%	98.14%
1940s	2754	2409	87.47%	98.18%
all years	7208	6242	86.60%	97.51%

Table 1: Identification rate and accuracy rate by decades. The number of applications includes only applications with a publication date.

When analysing the reasons for these differences, we examined whether the newspaper's structure and the placement of the text relative to the graphical representation affect them, as shown in

Figure 3. We were unable to find a regression discontinuity related to the structural changes. Therefore, we conclude that the quality of the scan, in conjunction with the condition of the paper, is the primary reason for the decrease in the identification rate of 20th-century papers. The absence of a structural discontinuity confirms that the declining identification rate in 1920s editions is a *form-level* effect, driven by physical paper degradation and scan quality rather than by layout changes.

Discussion

Historians, like other scholars, should recognise the limitations of the sources they rely on. For a while now, historians have been aware and conscious of the shortcomings of archives. As Luthra et al. (2024) put it: archives are "contested spaces of knowledge production". Thomas et al. (2017: 1) summarised this understanding: archives are not neutral. It matters who recorded what, how and when, which materials were preserved, and which are made accessible. No less important is the other side of the coin: Which materials were not collected to begin with, or not processed, archived and made available. The acknowledgement of these limitations has earned the title 'the silence of the archives'.

Archival silences are the result of archival power (Bruns, 2024: 12). Other than political censorship, national interests, and legal limitations such as copyright or privacy that may shape what is included and available in the archives and what is left outside, we point to a less explicit factor that shapes the contents of archives and hence may affect research. This is a technological bias, referring not only to the availability of relevant resources, but to their accessibility, readability, and hence, their usability. This bias focuses on the user's side rather than on the choices made by the archive's curators. The inherent affordances of the archival material may also influence the research, resulting in subtle biases.

One example of such technological bias is oral testimony, which is too often left outside the archives. A second example is of handwritten texts. Such texts may be included in the archives but nevertheless overlooked. It is often the case that the handwriting is unclear and difficult to decipher, and that, if not digitised, the document's quality may have deteriorated over the years. This affordance of handwriting renders such sources more difficult to explore. Some scholars may simply ignore them and focus on the printed texts, often the official ones. With the introduction of AI technologies, deciphering handwritten texts may become easier. Our personal experience in deciphering handwritten comments of British clerks in the 1920s-1940s indicated that most AI programs provided a relatively accurate transcription. In the meantime, the result is an inherent text bias: texts hold a higher place in the hierarchy.

However, not all printed texts are equal. Some texts are easier to read due to the condition of the original paper. Other inherent characteristics of the text also affect its usability. The case of Mandate Palestine's trademark data that we discussed in this article exemplifies this concern. Trademark data is textual and printed, accompanied by trademarks (some are text-only, some are graphical, some combine text and graphics, all usually in black and white), and we would expect them to be placed at a higher level in the hierarchy of resource availability. When we shift from manual to digital research, new, unexpected features become a challenge. As we discussed earlier, the older the text, the more difficult it is to achieve accurate machine reading. The layout of the official publications (one column, two columns, etc.) adds another difficulty, as do the changing practices of publication, e.g., the placement of the mark itself in relation to the accompanying text.

The concern is that these technological features would introduce a temporal bias, potentially skewing scholarly understanding toward more recent periods within a given historical timeframe.

Thus, the temporal bias we identified, where OCR accuracy improves for more recent publications within our historical corpus, reveals a systematic challenge that likely affects many digitisation projects dealing with materials spanning multiple decades. The 0.38 percentage points annual improvement in matching probability suggests that physical deterioration, evolving printing technologies, and paper quality create compounding effects that researchers must account for when interpreting digitised historical corpora. This finding challenges the assumption that digitised collections provide uniform access to historical periods and highlights the need for bias-aware methodologies in digital humanities research. Our findings align with and extend observations from other large-scale digitisation initiatives, such as those by Jarlbrink and Snickars (2017) and Suissa et al. (2022). However, the standardisation of the newspaper structure and the extended period allow us to quantify the effect in a more controlled way, to control for other sources of chronological bias, and to quantify its implications for historical research.

Addressing the technological-temporal bias is crucial. When examining the limits of the raw material and datasets, researchers engaging in big-data studies should be aware not only of archival silences, issues such as periodisation of the material found in archives, the scope of coverage, or limited availability due to political reasons, but also of inherent technological affordances. Thus, raising awareness is needed. The focus of this awareness is not about *what* is included or *who* is excluded, but about *how* the material is preserved. Attention shifts from the subject to the material. A second way to address the potential bias, once identified, is to devise better research tools that may equalise the quality of the raw material and render all on a similar technological level. A related tool is to measure accuracy levels, as we have suggested in this article.

The establishment of the British trademark registry in Palestine exemplifies how colonial administrative systems created both opportunities and challenges for historical research that resonate with contemporary concerns about colonial archives (Luthra et al., 2024; Candela et al., 2023). Unfortunately, unlike many other colonial registries (Decker, 2013), the original trademark registry did not survive; however, there were enough other sources, namely the official Gazette publications, to reconstruct it.

The standardised British model of 1875 for trademark registration was duplicated and imposed on imperial territories. This model provided a systematic framework for recording commercial activity, generating some of the documentary evidence that enabled reconstruction of economic and cultural life during the Mandate period; it also embedded colonial power structures within the archival record itself. The very existence of the registry was a governmental act; its contents were initiated by applicants but strictly dictated by the law as to which details should be included in the application and recorded in the trademark registry. There was no room for personal voice there, other than in the trademark itself.

The multilingual nature of trademark applications, as evident in records such as the Hassan Nabulsi case (Figure 1), reveals how local Palestinian, Arab, Jewish, and British merchants navigated and adapted colonial legal frameworks to their own commercial purposes. However, the systematic biases inherent in colonial documentation systems mean that the trademark registry, like other colonial archives, likely reflects the perspectives and priorities of the colonial administration rather than providing a balanced representation of all commercial actors. The over 300 gazette publications containing trademark notifications thus represent both a valuable repository of historically marginalised voices, local Palestinian and regional merchants whose commercial

activities might otherwise be lost to history, and a colonial artefact that requires careful critical analysis to avoid perpetuating the administrative silences and hierarchies embedded in the original British archival system.

Conclusions

In this study, we developed and evaluated a method for extracting distinctive information from poor-quality early twentieth-century newspapers. The 86.6% identification rate achieved demonstrates that automated extraction of structured data from deteriorated historical materials is not only feasible but can reach practical implementation thresholds when combined with appropriate technological approaches and methodological frameworks.

Some technical aspects of our method highlight the unique challenges of linking metadata from low-quality scans. Unlike many other digitisation projects (Drobac et al., 2017; Aske & Giardinetti, 2023), using machine-learning models for correcting spelling mistakes was not feasible in this case because the algorithm primarily focuses on extracting numbers, dates, names of individuals and corporations, countries, and cities – phrases that are not typically found in dictionaries and do not convey significant semantic information within sentences. This work evaluates solutions to overcome these challenges while achieving high accuracy in the extraction, identification, and matching of annotated graphical data.

Our findings suggest a chronological bias: the algorithm's efficiency degrades over time. This temporal factor has profound implications for the historical data gathered using this method. In particular, we demonstrate that our understanding of more recent periods, such as the 1940s, tends to be more comprehensive than that of older periods, for instance, the 1920s. The algorithm appears to match content more accurately for more recent periods, leading to a chronological bias in the information obtained. As Late & Kumpulainen (2022) note, understanding the roots and reasons behind digitisation quality is crucial for providing the basis for the documentation process.

Our methodology’s focus on maximising correct matches rather than perfect transcription represents a pragmatic approach that prioritises research utility over textual fidelity. This trade-off is particularly relevant for projects aimed at enabling large-scale analysis rather than producing diplomatic transcriptions, suggesting a broader methodological distinction that digital humanities projects should consider when defining their objectives and success metrics. The 86.6% success rate, while impressive in absolute terms, must be understood in the context of systematic temporal bias and the broader challenges of digitising historical materials, as documented across multiple large-scale projects spanning different languages, time periods, and document types.

Acknowledgement

We wish to thank Raz Ashkenazi for leading the legal research team. Birnhack acknowledges the support of ISF Grant 532/21.

References

ABBYY FineReader. Abbyy. https://www.abbyy.com/ (accessed March 20, 2022).

Ali, D., Milleville, K., Verstockt, S., Van de Weghe, N., Chambers, S., & Birkholz, J. M. (2024). Computer Vision and Machine Learning Approaches for Metadata Enrichment to Improve Searchability of Historical Newspaper Collections. Journal of Documentation, *80*(5), 1031-1056.

Arnold, T., Maples, S., Tilton, L., & Wexler, L. (2017). Uncovering Latent Metadata in the FSA-OWI Photographic Archive. *DHQ: Digital Humanities Quarterly*, *11*(2).

Aske, K., & Giardinetti, M. (2023). (Mis)matching Metadata: Improving Accessibility in Digital Visual Archives through the EyCon Project. *ACM Journal on Computing and Cultural Heritage*, *16*(4), 1-20.

Beebe, B. (2003). The Semiotic Analysis of Trademark Law. UCLA L. Rev., 51 621.

Bently, L. (2008). The Making of Modern Trade Mark Law: The Construction of the Legal Concept of Trade Mark (1860–1880). In: Bently, L., Davis, J., Ginsburg J. C., eds. Trade Marks and Brands: An Interdisciplinary Critique. Cambridge University Press; 3-41.

Bently, L. (2011). The "Extraordinary Multiplicity" of Intellectual Property Laws in the British Colonies in the Nineteenth Century. Theoretical Inquiries in Law. *12*(1): 161-200.

Bently, L. & Bone, R., eds. (2024). Research Handbook on the History of Trademark Law (Edward Elgar).

Birnhack, M. (2021). Colonial Trademark: Law and Nationality in Mandate Palestine, 1922–48. Law & Social Inquiry *46*(1): 192-225.

Birnhack, M. (2023). "Reconstructing the Trademark Registry of Mandate Palestine and What Historical Data Can Reveal". Trademark Reporter *113*: 815-837.

Bremer-Laamanen, M. (2006). The Present Past - the History of Newspaper Digitisation in Finland. IFLA Publications, 122, 43.

Bruns, E. (2024). Archival Silences and Avenues to Address Them: A Literature Review. The Serials Librarian, *85*(1-4): 11-18.

Candela, G., Pereda, J., Sáez, D., Escobar, P., Sánchez, A., Torres, A. V., Palacios, A.A., McDonough, K. & Murrieta-Flores, P. (2023). An Ontological Approach for Unlocking the Colonial Archive. ACM Journal on Computing and Cultural Heritage, *16*(4): 1-18.

Chiticariu, L., Li, Y., & Reiss, F. R. (2013). Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems! In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 827–832. Association for Computational Linguistics.

Cohen, A., Nissim, N., Rokach, L., and Elovici, Y. (2016). SFEM: Structural Feature Extraction Methodology for the Detection of Malicious Office Documents Using Machine Learning Methods. Expert Systems with Applications 63: 324-343.

Conway, P. (2013). Preserving Imperfection: Assessing the Incidence of Digital Imaging Error in HathiTrust. Preservation, Digital Technology & Culture, 42(1).

Corbett, M., Rao, M., and J. Teece, D. (2008). A Primer on Trademarks and Trademark Valuation." In Teece, D. J., ed. The Transfer and Licensing of Know-How and Intellectual Property: Understanding the Multinational Enterprise in the Modern World, 247-262. World Scientific.

Cordell, R. (2017). " Q i-jtb the Raven": Taking Dirty OCR Seriously. Book History, 20(1), 188-225.

Damerau, F.J. (1964). A Technique for Computer Detection and Correction of Spelling Errors. Communications of the ACM *7*(3): 171-176.

Decker, S. (2013). The Silence of the Archives: Business History, Post-colonialism and Archival Ethnography. Management & Organisational History, *8*(2): 155-173.

de la Rosa, J.D. (2025). Machine learning at the National Library of Norway. Navigating Artificial Intelligence for Cultural Heritage Organisations.

Drucker, J. (2020). Visualization and interpretation: Humanistic approaches to display. MIT Press.

Drobac, S., Kauppinen, P., and Lindén, K. (2017). OCR and Post-Correction of Historical Finnish Texts. In Proceedings of the 21st Nordic Conference on Computational Linguistics, 70-76.

Duguid, P. (2009). French Connections: The International Propagation of Trademarks in the Nineteenth Century. Enterprise and Society *10*(1): 3-37.

Duguid, P., da Silva L., and Mercer, J. (2010). Reading Registrations: An Overview of 100 Years of Trademark Registrations in France, the United Kingdom, and the United States. In: Lopes, da Silva, T. & Duguid, P. Trademarks, Brands, and Competitiveness 9-18. Routledge.

Dutta, A., Bergel, G., & Zisserman, A. (2021). Visual analysis of chapbooks printed in Scotland. In Proceedings of the 6th International Workshop on Historical Document Imaging and Processing (HIP '21). ACM. https://doi.org/10.1145/3476887.3476893

Ehrmann, M., Romanello, M., Clematide, S., Ströbel, P. B., & Barman, R. (2020). Language resources for historical newspapers: The Impresso Collection. In Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020) (pp. 958–968). ELRA.

Europeana Project. Europeana. (2022). https://www.europeana.eu/en/collections/topic/18-newspapers (accessed March 20, 2025).

Gooding, P. (2016). Exploring the Information Behaviour of Users of Welsh Newspapers Online through web log analysis. Journal of Documentation, *72*(2), 232-246.

Hill, M. J., & Hengchen, S. (2019). Quantifying the Impact of Dirty OCR on Historical Text Analysis: Eighteenth Century Collections Online as a Case Study. Digital Scholarship in the Humanities, *34*(4), 825-843.

Holley, R. (2009). How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. D-Lib Magazine 15(¾).

Jarlbrink, J., & Snickars, P. (2017). Cultural Heritage as Digital Noise: Nineteenth Century Newspapers in the Digital Archive. Journal of Documentation *73*(6): 1228-1243.

Kahle, P., Colutto, S., Hackl, G., & Mühlberger, G. (2017). Transkribus – A service platform for transcription, recognition and retrieval of historical documents. In Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition Workshops (pp. 19–24). IEEE.

Kiessling, B., Tissot, R., Stokes, P., & Stökl Ben Ezra, D. (2019). eScriptorium: An open source platform for historical document analysis. In Proceedings of the 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). IEEE. https://doi.org/10.1109/ICDARW.2019.10032

Landes, W. M., & Posner, R. A. (1988). The Economics of Trademark Law. Trademark Reporter, *78*: 267-270.

Late, E., & Kumpulainen, S. (2022). Open Image Collections–sources of Research Data?. Informaatiotutkimus, *41*(2–3): 88-91.

Likhovski, A. (2006). Law and Identity in Mandate Palestine. University of North Carolina Press.

Luthra, M., Todorov, K., Jeurgens, C., & Colavizza, G. (2024). Unsilencing colonial archives via automated entity recognition. Journal of Documentation, *80*(5): 1080-1105.

Neudecker, C., Baierer, K., Federbusch, M., Boenig, M., Würzner, K. M., Hartmann, V., & Herrmann, E. (2019). OCR-D: An end-to-end open source OCR framework for historical printed documents. In Proceedings of the 3rd international conference on digital access to textual cultural heritage (53-58).

Nockels, J., Gooding, P., & Terras, M. (2024). The implications of handwritten text recognition for accessing the past at scale. Journal of Documentation, *80*(7): 148-167.

Olson, L., & Berry, V. (2021). Digitisation decisions: comparing OCR software for librarian and archivist use. Code4Lib Journal, (52).

Philips, J., and Nasseh T. (2020). Historical Document Processing: A Survey of Techniques, Tools, and Trends. KDIR: 341-349.

Pletschacher, S., Clausner, C., & Antonacopoulos, A. (2015). Europeana Newspapers OCR Workflow Evaluation. Proceedings of the 3rd international workshop on historical document imaging and processing.

Rydberg-Cox, J. A. (2009). Digitising Latin incunabula: Challenges, methods, and possibilities. Digital Humanities Quarterly *3*(1).

Ruecker, S., Radzikowska, M., & Sinclair, S. (2016). Visual interface design for digital cultural heritage: A guide to rich-prospect browsing. Routledge.

Scardamaglia, A. (2015). Colonial Australian Trade Mark Law: Narratives in Lawmaking, People, Power and Place. Australian Scholarly.

Schechter, F. (1925). The Historical Foundations of the Law Relating to Trademarks. Columbia University Press.

Sherman, B. & Bently, L. (1999). The Making of Modern Intellectual Property Law. Cambridge University Press.

Smith, G. V., & Richey, S. M. (2013). Trademark valuation: A tool for brand management. John Wiley & Sons.

Smith, R. (2007). An overview of the Tesseract OCR engine. In IEEE Ninth international conference on document analysis and recognition (ICDAR) (Vol. 2, pp. 629-633).

Strange, C., McNamara, D., Wodak, J., & Wood, I. (2014). Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitised Historical Newspapers. DHQ: Digital Humanities Quarterly, 8(1).

Suissa, O., Zhitomirsky-Geffet, M., & Elmalech, A. (2022). Toward a Period-specific Optimised Neural Network for OCR Error Correction of Historical Hebrew Texts. ACM Journal on Computing and Cultural Heritage (JOCCH), *15*(2): 1-20.

Tafti, A.P., Baghaie, A., Assefi, M., Arabnia, H.R., Yu, Z., Peissig, P. (2016). OCR as a Service: An Experimental Evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym. In: Bebis, G., et al. Advances in Visual Computing. ISVC 2016. Lecture Notes in Computer Science, vol 10072. Springer, Cham.

Tanner, S., Muñoz, T., & Ros, P. H. (2009). Measuring Mass Text Digitisation Quality and Usefulness. D-lib Magazine, *15*(7/8): 1082-9873.

Turkel, W. J., Kee, K., & Roberts, S. (2012). A method for navigating the infinite archive. In History in the Digital Age (pp. 61-75). Routledge.

Thomas, D., Fowler, S., Johnson, V. (2017). The Silence of the Archive. Facet Publishing.

Van Strien, D., Beelen, K., Ardanuy, M. C., Hosseini, K., McGillivray, B., & Colavizza, G. (2020). Assessing the Impact of OCR Quality on Downstream NLP Tasks, ICAART 2020 - Proceedings of the 12th International Conference on Agents and Artificial Intelligence.

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information processing & management*, *50*(1), 104-112.

Whitelaw, M. (2015). 'Generous interfaces for digital cultural collections', Digital Humanities Quarterly 9(1).

Wilkins, M. (1992). "The Neglected Intangible Asset: The Influence of the Trade Mark on the Rise of the Modern Corporation", 34 Business History 66.

Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., & Zhou, M. (2020). LayoutLM: Pre-training of Text and Layout for Document Image Understanding. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1192–1200.