Beyond Archives: Designing a Segmentation and ATR Workflow for Mid-Twentieth-Century Italian Typescripts in a Relational Digital Library

Giulia Lembo[1]

[1]University of Naples "L'Orientale", Naples, Italy

Abstract

Beyond Archives is a digital-library project that connects archival science and digital humanities to model plural memory sources on the post-war Julian–Dalmatian exodus to Naples. This short paper presents an open, documented workflow for layout segmentation and Automatic Text Recognition (ATR) of mid-twentieth-century typescripts from the Prefecture of Naples. Implemented in eScriptorium with the Kraken engine, the workflow is embedded in a three-stage pipeline that runs from digitization through formalization to publication in a TEI- and IIIF-based relational digital library. It combines SegmOnto-based region and line classes, fine-tuned segmentation and ATR models, and explicit model cards and paradata released in an OCR/HTR repository. Evaluation on heterogeneous test sets shows layout and recognition quality adequate for downstream TEI encoding and named-entity extraction. The contribution speaks to discussions on FAIR-aware infrastructures for humanities-oriented digital libraries and on the role of document-analysis pipelines in supporting critical work with contested memory collections.

Keywords

automatic text recognition (ATR), layout segmentation, archival digital libraries, TEI-XML, Italian typescripts

1. Introduction

Archives and memory sources are never neutral [1]. They are shaped by power relations and by uneven capacities to record and transmit experience, especially in contexts of displacement and political conflict [2, 3]. Archives sit within wider memorial networks that include institutional records, community practices, informal collections, and oral transmission [4, 5]. Within this landscape, *Beyond Archives* brings together archival science and digital humanities to develop a relational digital library on the post-war Julian–Dalmatian exodus to Naples. The project treats institutional records and oral histories as parts of a single memorial ecosystem and asks how their entanglements can be represented, queried, and reused without erasing their tensions and asymmetries [6].

The project investigates how plural and contested memory sources can be digitized and archived while keeping conflicts visible and legible for different publics. The modular workflow moves from digitization to formalization and publication, applied in parallel to written and oral sources and grounded in open standards and FAIR principles (Findability, Accessibility, Interoperability, and Reuse)[7]. The written corpus consists of mid-twentieth-century typescripts from the Prefecture of Naples, produced between 1946 and 1958 to govern the reception of refugees. Digitized as IIIF[1] manifests in the *OrienTales* institutional repository of the University of Naples "L'Orientale", these records become the testbed for a domain-specific ATR pipeline. Building on work that reconceptualizes ATR as an umbrella term for Optical Character Recognition and Handwritten Text Recognition and stresses explicit model training and error analysis [8, 9], the project fine-tunes open models on degraded Italian typescripts and evaluates them at character and word level for downstream use in a relational digital-library setting. The resulting

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

*Corresponding author.

g.lembo2@unior.it (G. Lembo)

[1]International Image Interoperability Framework (IIIF): https://iiif.io

transcriptions support entity extraction, TEI[2] encoding, and critical exploration [10, 11].

The oral corpus, composed of semi-structured interviews with former refugees and community members, is recorded under ethical protocols (in line with the guidelines of the *Tavolo Permanente per le fonti orali*[3]), transcribed with speech-to-text tools, and revised and modeled alongside the written material in TEI-XML [12, 13]. Both streams converge in a TEI-based environment published through TEI Publisher[4] with IIIF integration, conceived as a relational space in which users can move between facsimiles, textual structures, and named entities. This paper focuses on the design and evaluation of the workflow for layout segmentation and ATR of mid-twentieth-century Italian typescripts, implemented in eScriptorium with the Kraken engine [15, 16] and conceived to support both the *Beyond Archives* case study and other small-scale collections of archival typescripts.

2. Background and related work

Debates in archival science and memory studies have long challenged the idea that archives are neutral containers and have emphasized their role as historically situated institutions that organize and authorize specific versions of the past [17, 3, 1, 18, 19]. Archives materialize distributions of power and capacity to record, preserve, and transmit experience, particularly in situations of displacement and political violence where institutional and community ways of classifying and narrating events are deeply misaligned [2, 20]. Approaches such as travelling and multidirectional memory foreground movement and negotiation in the circulation of memories, inviting us to see testimonies, institutional records, and commemorative practices as dynamically entangled layers of the same memorial field [21, 22, 23]. Building on these insights, archival scholarship has challenged metaphorical uses of "the Archive" and stressed the historically situated character of records, provenance, and evidential value, tied to institutional histories, administrative reforms, and political choices [24]. The Australian records-continuum model reframes recordkeeping as an ongoing, multi-layered process across institutions, individuals, and communities [25, 26] and resonates with the notion of an archival multiverse where heterogeneous recordkeeping traditions and communities of records coexist and assert their own rights to memory [6, 27, 5].

Digital-humanities and knowledge-organization studies foreground modeling, formalization, and the design of access systems. They show that humanities data must be expressed in explicit, shared structures to become computable and can enhance archival sources by making structures, relationships, and gaps more visible [12, 13]. Knowledge-organization work stresses that any system rests on theoretical and social assumptions [28, 29]. In the archival field, these concerns converge in work that highlights how standards and descriptive practices embed particular conceptions of provenance, context, and reuse and can foreground some users while marginalizing others [30, 31]. As digital reuse spreads, records and descriptive surrogates circulate across platforms, are reaggregated into new collections, and acquire meanings that depend on emergent contexts of use [32, 33, 34, 35, 36]. Documenting how records are transformed, re-described, and linked becomes part of archival responsibility, especially when they concern vulnerable communities and contested events.

Within this broader landscape, recent work on ATR and document analysis adds a methodological layer that is central to *Beyond Archives*. Chiffoleau [9] proposes to treat ATR as an umbrella term encompassing both OCR and HTR and to understand model training, ground truth, and prediction errors as components of a documented process rather than an opaque preprocessing step. Experiments with large, multilingual base models such as CATMuS-Print [Large] show that such models can offer a strong starting point but must be adapted to specific languages, typefaces, and archival practices to reach quality levels adequate for humanities-oriented tasks [10, 11]. In the Italian context, applications of automatic text recognition to archival heritage have demonstrated that HTR can be integrated into

[2] Text Encoding Initiative (TEI): https://tei-c.org

[3] The *Tavolo Permanente per le fonti orali* is a national working group that brings together scholars, archivists, and institutions involved in oral history. https://sites.google.com/view/tavolopermanenteperlefontioral/home.

[4] TEI Publisher: https://teipublisher.com; see also: [14]

historical and documentary workflows and that its value lies in the combination of increased speed with new forms of access and analysis enabled by careful modeling and domain expertise [37, 38]. At the level of evaluation, Bubula et al. [8] argue that quality assessment must grapple with heterogeneous documents and with the limits of small ground-truth samples, combining aggregate metrics with qualitative analysis. *Beyond Archives* adopts and extends these insights by treating model-training logs, ground truth, and evaluation artifacts as part of the documented infrastructure of the project and by aligning segmentation and ATR with archival reflections on reuse and the transformation of context [4, 24, 31].

3. Case study and corpora

The case study focuses on the post-war Julian–Dalmatian exodus viewed from Naples. Several thousand refugees from Istria, Fiume, and Dalmatia were housed in temporary camps and collective facilities in the city [39]. This local configuration offers a dense observatory for studying how plural memories are constructed, transmitted, and contested across institutional and community settings [40]. Within *Beyond Archives*, Naples functions as a laboratory where institutional records and oral recollections can be brought into relation without dissolving their differences.

On the institutional side, the project concentrates on the records of the Prefecture of Naples concerning refugee management between 1946 and 1958. A first corpus has been identified in six boxes within the series "Prefettura di Napoli – III versamento, categoria V 'Varie'," inventoried in the *INVENTARIO Fondo Prefettura di Napoli* edited by Giuliana Buonaurio. The files include administrative correspondence, nominal lists, circulars, and reports and consist mainly of mid-twentieth-century Italian typescripts on fragile paper, often with stamps, marginal notes, and corrections. Their relatively regular layouts, combined with overlaps between typed text and rubber stamps and with variable legibility, make them a suitable testbed for segmentation and ATR. In the wider project they form the written backbone of the relational digital library, where they are linked to places, people, and institutions and queried alongside oral sources.

On the community side, the project is building a corpus of semi-structured interviews with former refugees, relatives, and other witnesses. The interviews are conducted under ethical protocols that include informed consent and negotiated levels of access, are recorded and transcribed with speech-to-text tools, and are revised for research use. While the present contribution does not discuss the oral corpus in detail, the segmentation and ATR workflow is developed with this dual horizon in mind, with written and oral components conceived as parallel streams that will be modeled in TEI and interlinked in the same digital library.

4. Workflow design for document analysis

Within the broader *Beyond Archives* project, this short paper focuses on the workflow for the written corpus, showing how digitized images of Prefecture records are transformed into machine-readable text and prepared for TEI modeling and publication. Digitization produces high-quality images and IIIF manifests that anchor written records in a stable, citable environment. Formalization transforms these materials into textual and semantic representations through layout segmentation, ATR, semi-automatic entity extraction, and TEI-based modeling. Publication exposes the resulting representations in a relational digital library implemented in TEI Publisher, an open-source framework for publishing TEI collections on top of the native XML database eXist-db, with IIIF integration. Across these stages, the project seeks to keep data Findable, Accessible, Interoperable, and Reusable in line with FAIR principles.

Archival records move through eScriptorium and Kraken [15, 16], while interviews are processed through speech-to-text tools (e.g. OpenAI's Whisper [41]) and prepared for encoding with analogous principles of cleaning and minimal alignment. The present contribution concentrates on the document-analysis workflow for the written sources, which forms the backbone of the formalization stage and provides the testbed for the segmentation and ATR models discussed in this paper.

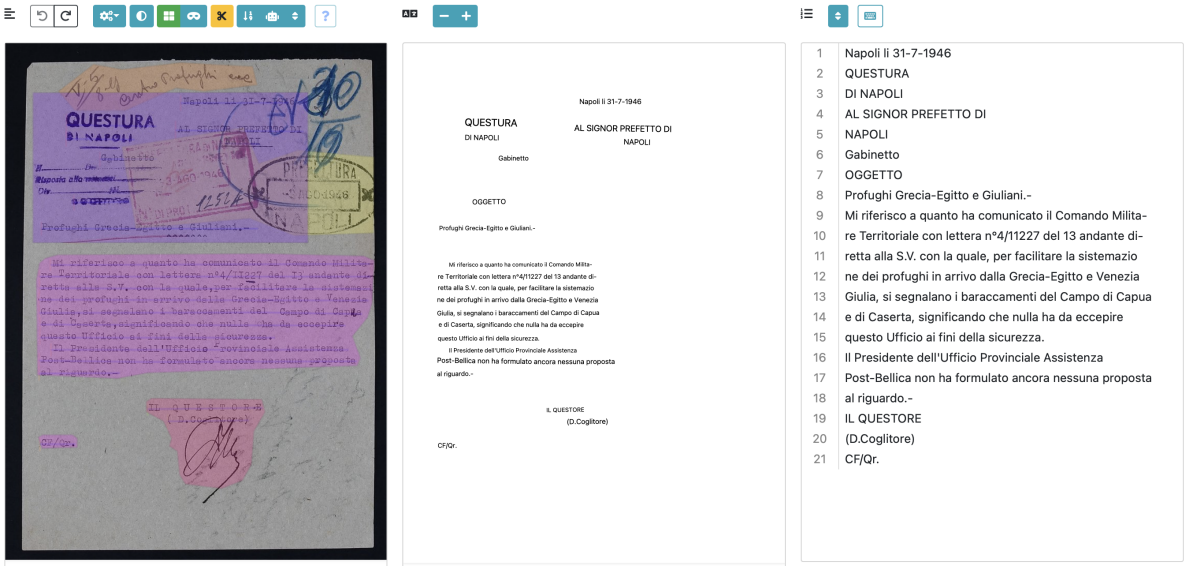

Figure 1: Example Prefecture page in eScriptorium, showing the scanned document (left), normalized layout (centre), and line-level transcription (right).

4.1. Overall architecture and data flow

Within the three-stage workflow, document analysis occupies the space between image creation and high-level modeling. It operates on digitized records from the Prefecture corpus that have been ingested into the institutional repository as IIIF manifests. For each selected file, the IIIF image is imported into eScriptorium, where region and line segmentation are produced, refined, and stored. The segmented images then feed into ATR models fine-tuned on similar typescripts. The resulting transcriptions are exported in PAGE XML and ALTO XML, which act as master representations of layout and text.

This PAGE XML layer connects document analysis to the modeling and publication environment. PAGE XML files are mapped to TEI-XML by aligning regions and lines with TEI structures such as divisions, paragraphs, and notes, following a broader understanding of formalization that treats markup and modeling as central to humanities data work [12, 13]. The TEI files are ingested into TEI Publisher, linked back to IIIF facsimiles, and exposed through search and browsing interfaces. Segmentation and ATR thus form the core of the formalization stage in a continuous pipeline that starts from digitized archival records and ends in a navigable, queryable digital library.

4.2. Segmentation workflow

The segmentation component uses SegmOnto [42] as a starting point for defining region and line classes and adapts its vocabulary to the specific needs of the Prefecture corpus. In addition to generic categories for main text and headings, the project distinguishes zones such as headings, body text, signatures, stamps, and marginal notes. These classes reflect the documentary reality of the corpus, where many pages contain a mixture of formal textual content, rubber stamps, signatures, and later archival notes that must be distinguished if the resulting TEI representation is to preserve provenance and context. Figure 1 illustrates a typical page from the corpus as processed in eScriptorium, with the scanned document, normalized layout, and line-level transcription.

Training data for segmentation were created in eScriptorium by manually annotating a representative set of pages drawn from different fascicles and document types. The annotated pages were split into training, validation, and test subsets, with the split recorded in explicit lists and metadata files. A first segmentation model, SegModel-A, was trained on an initial pool of pages and evaluated on a held-out set A that functions as the official test set, then examined qualitatively on additional pages not used in training to identify typical errors and underrepresented layouts. The training set was subsequently enriched with pages featuring complex layouts, heavy stamps, or marginal notes, and a second model,

SegModel-B, was trained on the expanded pool and evaluated on a new internal set A and on a second evaluation set composed of diagnostically chosen pages. In practical terms, the segmentation workflow is designed so that new pages and categories can be incorporated iteratively and so that training and evaluation splits remain explicitly documented for each model version [8]. At this stage evaluation relies on qualitative manual scoring; in future iterations we plan to complement it with standard segmentation metrics such as Document Layout Error Rate (DLER)[43] so that our internal checks remain comparable with other document-analysis workflows.

4.3. ATR workflow

The ATR component is built on top of large, multilingual base models and is tailored to the language, typography, and degradation patterns of the Prefecture corpus. Following the perspective that understands ATR as a continuum encompassing both OCR and HTR [9], the project starts from a general model trained on printed and typewritten sources and then fine-tunes it on curated subsets of Italian administrative typescripts. The training material consists of page-level ground truth produced in eScriptorium, where lines are corrected manually to reflect the original content while preserving layout information in PAGE XML.

Ground-truth pages are grouped into subsets that correspond to different document batches in the corpus. Each subset is split into training, validation, and test partitions with explicit lists that record which pages belong to which split, typically with around forty training pages and small validation and test sets. These splits are used consistently across successive fine-tuning runs so that improvements or regressions can be attributed to changes in data or configuration rather than to variations in the evaluation sample. A base model is first adapted to one subset of Prefecture documents and then further adapted to additional subsets that differ slightly in layout or typing conventions, such as lists or narrative reports, in order to obtain a domain-specific model that can generalize across the corpus. Throughout this sequence, the project records training parameters, validation curves, and test-set results for each run, in line with the idea that ATR should be treated as a documented process [9]. Within the triad of digitization, formalization, and publication, this ATR workflow turns visual records into textual data that can be modeled, encoded, and reused.

4.4. Documentation, paradata, and publication of models

A central design choice of the workflow is to treat segmentation and ATR artifacts as part of the scholarly record. For each model version, the project maintains a bundle that includes the trained model, the lists defining training, validation, and test splits, a selection of representative ground-truth pages, and a short textual description of the training configuration and typical error patterns. These bundles support both internal interpretation and external reuse and document how models were obtained and under which conditions their reported error rates should be understood. At the level of TEI encoding, references to the models and their versions are recorded in the `teiHeader`, together with notes about ground-truth creation, correction policies, and post-processing rules [30, 12, 13]. By making links between TEI files, segmentation models, ATR models, and source images explicit, the project enables users and future curators to reconstruct how a given transcription was produced and to evaluate whether it is adequate for their purposes. In line with the commitment to reuse expressed in the broader project design, the fine-tuned segmentation and ATR models, together with a curated subset of ground truth and the core configuration files, will be deposited in the OCR and HTR model repository on Zenodo as a shareable research object that can be cited, inspected, and adapted beyond the local project context, strengthening the FAIR profile of the overall infrastructure.

5. Early findings and evaluation

The first group of early findings concerns layout analysis. Table 1 summarizes the ground-truth lots and the segmentation and ATR models discussed in this section. For segmentation, a manual protocol scores

Table 1
Ground-truth lots and associated segmentation and ATR models.

Name	Type	Description
Lot 1	Ground-truth pages	52 pages (segmentation and ATR ground truth)
Lot 2	Ground-truth pages	50 pages (segmentation and ATR ground truth)
Lot 3	Ground-truth pages	Lots 1 and 2 merged (102 pages)
Lot 4	Ground-truth pages	50 pages (segmentation ground truth only)
SegModel-A	Segmentation model	Trained on Lots 1 and 2
SegModel-B	Segmentation model	Trained on Lots 1, 2, and 4
ATRModel-1	ATR model	Trained on Lot 1
ATRModel-2	ATR model	Trained on Lot 2, starting from ATRModel-1
ATRModel-3	ATR model	Trained on Lot 3, starting from ATRModel-2

Table 2
Summary of ATR evaluation on Prefecture test splits

Model	Training lot(s)	Test split	CER baseline (%)	CER fine-tuned (%)	WER baseline (%)	WER fine-tuned (%)
ATRModel-1	Lot 1 (52 pages)	Lot 1 test split	9.28	3.35	33.14	13.01
ATRModel-2	Lot 2 (50 pages)	Lot 2 test split	9.85	4.91	31.96	18.29
ATRModel-3	Lot 3 (102 pages)	Lot 3 test split	10.61	4.80	36.83	17.72

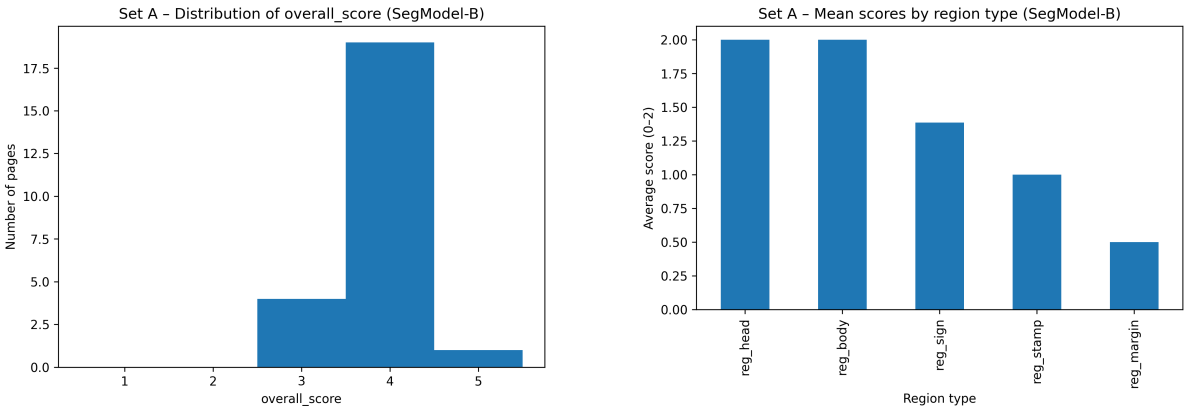

Figure 2: Segmentation results for SegModel-B on internal set A: (a) histogram of overall page scores; (b) mean scores per region.

each page on a 0–2 scale for regions and line quality, aggregates these scores into an overall 0–5 page score, and summarizes results by layout type. In the first fine-tuning cycle, SegModel-A was trained on about one hundred manually segmented pages from Lots 1 and 2 and evaluated on an internal test set A and on an external set B1. In the second cycle, SegModel-B was trained on an expanded pool of around one hundred and fifty pages that also included Lot 4 and was evaluated on a new internal set A built on its own split. On the internal sets A, both models perform well on the core regions of the documents. With the richer ground truth that includes Lot 4, SegModel-B yields more balanced mean scores, with head and body reaching the maximum value on the 0–2 scale and improved behavior on signatures and stamps. Overall scores concentrate in the upper part of the 0–5 scale and pages with low scores are reduced. Figure 2(a) summarizes the distribution of overall page scores, while Figure 2(b) reports mean scores per region for SegModel-B on its internal set A. The second group of findings concerns the ATR models fine-tuned in Kraken. The training strategy follows an incremental design, with three models (ATRModel-1, ATRModel-2, ATRModel-3) trained on successive lots of Prefecture ground truth and on

a combined dataset. Table 2 summarizes the most relevant results on the official test splits. Across the configurations, fine-tuning reduces character error rate from about 9–11 percent to roughly 3–5 percent and word error rate from about 32–37 percent to roughly 13–18 percent. Internal monitoring tables (key performance indicators, KPIs) confirm that character accuracy reaches values around 96–97 percent, while word accuracy remains above 80 percent. In the current workflow, ATR output is systematically revised before publication, so these metrics are used to estimate correction effort, prioritize additional fine-tuning, and inform lightweight post-processing. Qualitative error profiles show that residual errors cluster around whitespace mismatches and recurrent letter substitutions, which can be further mitigated through targeted correction rules. Under these conditions, the fine-tuned models provide a solid basis for downstream named-entity recognition and TEI encoding.

6. Discussion and conclusions

The results above indicate that the proposed workflow offers a coherent framework for segmenting and recognizing mid-twentieth-century Italian typescripts within a broader digital-library project. By making model training and evaluation explicitly documented, the workflow turns segmentation and ATR into accountable components of the infrastructure, where technical choices can be inspected and related to downstream uses. The combination of curated test sets, manual region scores, and standard ATR metrics produces a layered picture of model behavior that can be read in relation to TEI encoding, named-entity recognition, and interface design and revisited when the project evolves or when other researchers consider reusing the models, in line with archival reflections on reuse and the transformation of context [31]. The current workflow still presents limitations. The experiments focus on a specific archival series in a single language and time frame, so performance on other Italian administrative fonds or on different periods remains to be tested. Stamps, marginal notes, and heavily degraded pages are problematic for both segmentation and recognition, and accuracy is expected to degrade on the most challenging pages, where noise and overlaps are frequent. Methodologically, the 0–2 and 0–5 scoring scales have proven useful to structure qualitative judgments, but they remain manual and subjective and do not yet incorporate user-oriented evaluation of how segmentation and recognition errors affect search and exploration in the digital-library interface.

Future work will extend the training data for segmentation, with a specific focus on layouts where stamps, marginalia, and complex letterheads are prominent, and will test whether incorporating these pages improves the stability of region boundaries and line order without overfitting. A similar strategy will be explored for ATR, enlarging the training corpus to include other archival series and monitoring how error rates evolve across different subsets. On the project side, the document-analysis workflow will be more tightly connected to the oral sources of *Beyond Archives*, aligning transcripts and media files, applying shared entity-recognition pipelines to written and oral corpora, and exposing integrated views in TEI Publisher that allow users to move between administrative records and testimonies along people, places, and institutional actors. In this sense, the project takes seriously the idea of modeling as a way of making structures, relationships, and contexts explicit in a form that is simultaneously human-readable and machine-actionable [13]. Openness and reuse are central to this trajectory. The trained segmentation and recognition models, together with their ground truth and evaluation logs, are intended to be released in an OCR and HTR model repository on Zenodo, with clear paradata and licensing, so that other projects working on mid-twentieth-century Italian typescripts or on analogous administrative archives can test and adapt them. By coupling model release with transparent documentation of training and evaluation choices, the workflow aspires to make document analysis for contested-memory collections more portable and more open to critique, so that subsequent uses can build on it, question it, and extend it in directions that respond to other archival and community contexts. In the wider debate on the science and foundations of modern humanities libraries, the contribution lies in showing how document-analysis pipelines can be designed as FAIR-aware, reusable infrastructures and how their artifacts can be treated as part of the scholarly record, tailored to a specific archival corpus yet intended to circulate beyond it.

References

[1] M. Hedstrom, Archives and Collective Memory: More than a Metaphor, Less than an Analogy, in: T. Eastwood, H. MacNeil (Eds.), *Currents of Archival Thinking*, Libraries Unlimited, Santa Barbara, CA and Denver, CO, 2010, pp. 163–179.

[2] J. A. Bastian, *Owning Memory: How a Caribbean Community Lost Its Archives and Found Its History*, number 99 in Contributions in Librarianship and Information Science, Libraries Unlimited, Westport, CT, 2003. doi:10.5040/9798400694752.

[3] C. Pavone, Archivi e Potere, *Contemporanea* 12 (2009) 211–217.

[4] M. Hedstrom, Archives, Memory, and Interfaces with the Past, *Archival Science* 2 (2002) 21–43. doi:10.1007/BF02435629.

[5] A. J. Gilliland, Archival and Recordkeeping Traditions in the Multiverse and Their Importance for Researching Situations and Situating Research, in: A. J. Gilliland, S. McKemmish, A. J. Lau (Eds.), *Research in the Archival Multiverse*, Monash University Publishing, 2017, pp. 31–73. doi:10.26530/oapen_628143.

[6] S. McKemmish, M. Piggott, Toward the Archival Multiverse: Challenging the Binary Opposition of the Personal and Corporate Archive in Modern Archival Theory and Practice, *Archivaria* 76 (2013) 111–114.

[7] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, et al., The FAIR Guiding Principles for Scientific Data Management and Stewardship, *Scientific Data* 3 (2016) 160018. doi:10.1038/sdata.2016.18.

[8] M. Bubula, K. Baierer, J. Lehmann, C. Neudecker, V. Rezanezhad, D. Škarić, How Scalable Is Quality Assessment of Text Recognition? A Combination of Ground Truth and Confidence Scores, *Anthology of Computers and the Humanities* 3 (2025) 1286–1310. doi:10.63744/GR59c1iXu6Wj.

[9] F. Chiffoleau, Understanding the Automatic Text Recognition Process: Model Training, Ground Truth and Prediction Errors, 2025. Le Mans Université and Inria Paris.

[10] P. B. Ströbel, S. Clematide, M. Volk, T. Hodel, Transformer-Based HTR for Historical Documents, *arXiv* (2022). doi:10.48550/ARXIV.2203.11008, version 1, preprint. Paper presented at ComHum 2022, Lausanne.

[11] S. Gabay, T. Clérice, CATMuS-Print [Large], Preprint, Zenodo, 2024. doi:10.5281/zenodo.10592716.

[12] F. Tomasi, D. Buzzetti, *Metodologie informatiche e discipline umanistiche*, Manuali Universitari. Linguistica, 3 ed., Carocci Editore, Roma, 2012.

[13] F. Tomasi, *Organizzare la conoscenza: Digital Humanities e Web semantico*, Editrice Bibliografica, Milano, 2022. doi:10.53134/9788893573573.

[14] H. Scheithauer, L. Romary, Which TEI Representation for the Output of Automatic Transcriptions and Their Metadata? An Illustrated Proposition, *Journal of the Text Encoding Initiative* (2024). doi:10.4000/13e5a.

[15] B. Kiessling, Kraken – A Universal Text Recognizer for the Humanities, DataverseNL, application/pdf, 2019. doi:10.34894/Z9G2EX.

[16] B. Kiessling, R. Tissot, P. Stokes, D. Stokl Ben Ezra, eScriptorium: An Open Source Platform for Historical Document Analysis, in: *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2019, pp. 19–19. doi:10.1109/ICDARW.2019.10032.

[17] L. Giuva, S. Vitali, I. Zanni Rosiello, *Il potere degli archivi: usi del passato e difesa dei diritti nella società contemporanea*, B. Mondadori, Milano, 2007.

[18] Z. Lian, A History of Archival Ideas and Practice in China, in: A. J. Gilliland, S. McKemmish, A. J. Lau (Eds.), *Research in the Archival Multiverse*, Monash University Publishing, 2017, pp. 96–121. doi:10.26530/oapen_628143.

[19] G. Di Marcantonio, E se l'archivio non rispecchia l'istituto? Pavone e il rispecchiamento: analisi di una bozza preliminare, *AIDAinformazioni* 40 (2022) 51–68. doi:10.57574/596516313.

[20] M. Caswell, Seeing Yourself in History, *The Public Historian* 36 (2014) 26–37. doi:10.1525/tph.2014.36.4.26.

[21] M. Rothberg, *Multidirectional Memory: Remembering the Holocaust in the Age of Decolonization*, Cultural Memory in the Present, Stanford University Press, Stanford, CA, 2009.
[22] A. Erll, Travelling Memory, *Parallax* 17 (2011) 4–18. doi:10.1080/13534645.2011.605570.
[23] C. De Cesari, A. Rigney (Eds.), *Transnational Memory: Circulation, Articulation, Scales*, De Gruyter, Berlin, 2014. doi:10.1515/9783110359107.
[24] M. Caswell, 'The Archive' Is Not an Archives: Acknowledging the Intellectual Contributions of Archival Studies, *Reconstruction* 16 (2016) 21–36.
[25] S. McKemmish, F. H. Upward, B. Reed, The Records Continuum Model, in: *Encyclopedia of Library and Information Sciences*, 3 ed., Taylor and Francis, New York, 2009.
[26] S. McKemmish, Recordkeeping in the Continuum, in: A. J. Gilliland, S. McKemmish, A. J. Lau (Eds.), *Research in the Archival Multiverse*, Monash University Publishing, 2017, pp. 122–160. doi:10.26530/oapen_628143.
[27] A. Gilliland, Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment, in: M. V. Cloonan (Ed.), *Preserving Our Heritage: Perspectives from Antiquity to the Digital Age*, Neal-Schuman, ALA Editions, Chicago, 2015, pp. 150–161.
[28] C. Gnoli, V. Marino, L. Rosati, *Organizzare la conoscenza: dalle biblioteche all'architettura dell'informazione per il web*, Hops+Tecniche Nuove, Milano, 2006.
[29] B. Hjørland, What Is Knowledge Organization (KO)?, *Knowledge Organization* 35 (2008) 86–101. doi:10.5771/0943-7444-2008-2-3-86.
[30] G. Michetti, Unneutrality in Archival Standards and Processes, 2015. doi:10.5281/zenodo.16421, complete version. Abridged version published in F. Pehar, C. Schlögl, and C. Wolff, eds., *Re:inventing Information Science in the Networked Society. Proceedings of the 14th International Symposium on Information Science (ISI 2015)*, Zadar, 19–21 May 2015, 144–159. Glückstadt: Verlag Werner Hülsbusch.
[31] G. Michetti, Il riuso come trasformazione del contesto. Una prospettiva archivistica, *Digitalia* 18 (2023) 99–112. doi:10.36181/digitalia-00078.
[32] C. L. Borgman, What are digital libraries? Competing visions, *Information Processing & Management* 35 (1999) 227–243. doi:10.1016/S0306-4573(98)00059-4.
[33] A. M. Tammaro, Che cos'è una biblioteca digitale?, *Digitalia* 0 (2005) 14–33. URL: https://air.unipr.it/handle/11381/1511240.
[34] A. M. Tammaro, Biblioteca digitale e scienze umane. Parte II: La Biblioteca digitale di ricerca per l'apprendimento, 2008. URL: https://air.unipr.it/handle/11381/1895225.
[35] C. Damiani, Nomina nuda tenemus. Riformulare il senso archivistico, in: L. Pezzica, F. Valacchi (Eds.), *Dimensioni archivistiche: una piramide rovesciata*, Editrice Bibliografica, Milano, 2021.
[36] F. Valacchi, *L'archivio aumentato: tempi e modi di una digitalizzazione critica*, volume 9 of *In archivio*, Editrice Bibliografica, Milano, 2024.
[37] V. I. Schwarz-Ricci, Handwritten Text Recognition per registri notarili (secc. XV–XVI): una sperimentazione, *Umanistica Digitale* (2022) 171–181. doi:10.6092/ISSN.2532-8816/14926.
[38] S. Spina, Historical Network Analysis & HTR Tool. Per un approccio storico metodologico digitale all'archivio Biscari di Catania, *Umanistica Digitale* (2023) 163–181. doi:10.6092/ISSN.2532-8816/15159.
[39] D. Lazzarich, Foiba: Genealogy of an Untranslatable Word, in: S. Radstone, R. Wilson (Eds.), *Translating Worlds: Migration, Memory, and Culture. Creative, Social, and Transnational Perspectives on Translation*, Routledge, London, 2021, pp. 101–117.
[40] S. Radstone, R. Wilson (Eds.), *Translating Worlds: Migration, Memory, and Culture. Creative, Social, and Transnational Perspectives on Translation*, Routledge, London, 2021.
[41] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust Speech Recognition via Large-Scale Weak Supervision, *arXiv* (2022). doi:10.48550/ARXIV.2212.04356, version 1, preprint.
[42] S. Gabay, J.-B. Camps, A. Pinche, C. Jahan, SegmOnto: Common Vocabulary and Practices for Analysing the Layout of Manuscripts (and More), in: 1st International Workshop on Computational Paleography (IWCP@ICDAR 2021), Lausanne, 2021. URL: https://hal.science/hal-03336528.

[43] A. Vesalainen, M. Tolonen, L. Ruotsalainen, Document Layout Error Rate (DLER) Metric to Evaluate Image Segmentation Methods, *Machine Learning with Applications* 18 (2024) 100606. doi:10.1016/j.mlwa.2024.100606.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT and Grammarly in order to perform text translation, grammar and spelling check, and rephrasing. After using these tools/services, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.