15º Congresso BAD

PAPER

Aplicações da Inteligência Artificial na leitura e acesso à informação das papeletas médicas (1910-1926) dos Hospitais da Universidade de Coimbra

Diéssica Braga-Loth[a], Vítor Matos[a,b], Ana Margarida Dias da Silva[c]

[a]CIAS – Centro de Investigação em Antropologia e Saúde, Departamento de Ciências da Vida, Universidade de Coimbra, Portugal, diessica.bsilva@gmail.com
[b]Laboratório de Antropologia Biológica, Departamento de Biologia, Escola de Ciências e Tecnologia, Universidade de Évora, Portugal, vitor.matos@uevora.pt
[c]CHSC – Centro de História da Sociedade e da Cultura, Departamento de Ciências da Vida, Universidade de Coimbra, Portugal, anamargarida.silva@uc.pt

Resumo

A Paleografia tem sido recentemente revitalizada pelo desenvolvimento de tecnologias digitais e de Inteligência Artificial, que permitem automatizar processos de leitura de manuscritos e ampliar o acesso a conteúdos históricos. Inserido nesse contexto, o presente trabalho aplica métodos de Reconhecimento de Texto Manuscrito à transcrição das papeletas médicas (1910-1936) dos Hospitais da Universidade de Coimbra, preservadas no Arquivo da Universidade de Coimbra. Os testes iniciais com modelos genéricos de Reconhecimento de Texto Manuscrito em português, disponíveis na plataforma Transkribus, revelaram resultados insatisfatórios devido à coexistência, nas papeletas, de escrita datilografada e manuscrita e à diversidade caligráfica. Estes desafios demonstram a necessidade de desenvolver um modelo específico, adaptado à natureza e terminologia médica destes registos. O projeto alia Paleografia, Antropologia e Ciência da Informação, mostrando que a automatização só é eficaz quando acompanhada por curadoria e validação humanas. A Inteligência Artificial é aqui entendida como uma ferramenta colaborativa que amplia o olhar do investigador e contribui para a preservação ativa do património documental e científico, reforçando o papel dos arquivos como espaços de produção e difusão de conhecimento.

Palavras-chave: *Handwritten Text Recognition* (HTR), Inteligência Artificial, Paleografia digital, Papeletas médicas, *Transkribus*.

Introdução

A transformação digital das últimas décadas alterou profundamente as formas de produção, preservação e acesso à informação, com consequências diretas no trabalho desenvolvido em arquivos, bibliotecas e outras instituições de memória. No campo das Ciências Sociais e Humanas, este processo impulsionou o desenvolvimento das Humanidades Digitais, entendidas como um espaço interdisciplinar que integra métodos computacionais à investigação tradicional sem abandonar os princípios críticos e hermenêuticos das disciplinas históricas (Dacos, 2011; Kirschenbaum, 2010). Neste contexto, a digitalização massiva de documentos e coleções, a criação de repositórios digitais e o desenvolvimento

https://doi.org/10.48798/congressobad.3186

de ferramentas de análise automatizada passaram a ampliar significativamente as possibilidades de acesso, organização e reutilização do património documental (Ackel, 2021; Silva, 2025).

A crescente circulação digital da informação modificou igualmente a própria relação entre investigador, documento e arquivo. Como observa Le Coadic (1996), a informação não corresponde apenas a um dado objetivo, mas integra processos de comunicação, interpretação e construção de significado. Também Cornelius (2005) sublinha que o conteúdo informacional depende do contexto cultural e institucional em que é produzido e interpretado, evidenciando que os documentos não constituem reflexos neutros da realidade. Neste sentido, os arquivos devem ser compreendidos não apenas como espaços de preservação, mas como sistemas ativos de produção e organização do conhecimento (Zeitlyn, 2012). Os arquivos resultam de práticas institucionais específicas, refletindo critérios de seleção, formas de classificação e relações de poder que influenciam aquilo que é preservado, descrito e disponibilizado para investigação (Borges et al., 2023)

No domínio historiográfico, esta ampliação conceptual das fontes documentais acompanha transformações iniciadas ao longo do século XX, quando o conceito de fonte histórica deixou de se restringir exclusivamente aos textos oficiais escritos. A valorização de novos tipos de fontes e a incorporação de abordagens interdisciplinares expandiram significativamente as possibilidades de investigação histórica, permitindo integrar vestígios materiais, documentação administrativa, produções orais e visuais, bem como registos produzidos por diferentes instituições sociais (Le Goff, 1990; Barros, 2019). Esse alargamento metodológico contribuiu também para incluir comunidades e indivíduos frequentemente marginalizados nas narrativas históricas (Sanjad, 2017). Paralelamente, os avanços tecnológicos passaram a ocupar um papel central na preservação e análise documental, possibilitando não apenas a digitalização de acervos, mas também a aplicação de ferramentas computacionais à leitura e interpretação de manuscritos históricos (Lose et al., 2024).

É neste cenário que a Paleografia adquire nova relevância no contexto digital. Tradicionalmente dedicada ao estudo histórico das escritas e à interpretação de manuscritos, a disciplina tem sido progressivamente transformada pelas Humanidades Digitais, incorporando métodos computacionais capazes de ampliar a leitura, a transcrição e a análise de documentação histórica (Ackel, 2021; Ciula, 2017). A designada Paleografia Digital combina práticas filológicas, arquivísticas e tecnológicas, promovendo uma aproximação entre leitura humana e processamento algorítmico. Assim, a análise paleográfica deixa de se limitar à interpretação do manuscrito por um técnico especialista e passa a integrar sistemas de reconhecimento automático, bases de dados e ambientes colaborativos de transcrição, sem abandonar a necessidade de contextualização histórica e crítica documental (Borges & Silva, 2018; Ciula, 2017).

O desenvolvimento da Paleografia Digital está diretamente associado ao avanço de tecnologias de Inteligência Artificial (IA) aplicadas à leitura e processamento de manuscritos históricos. Nas últimas décadas, ferramentas de Reconhecimento de Texto Manuscrito (*Handwritten Text Recognition* – HTR) passaram a ocupar um papel central nas Humanidades Digitais, permitindo automatizar parcialmente processos tradicionalmente dependentes da leitura paleográfica por técnico especialista (Muehlberger et al., 2019; Nockels et al., 2022). Estas tecnologias utilizam modelos de aprendizagem automática treinados a partir de imagens e transcrições alinhadas, possibilitando o reconhecimento de padrões gráficos específicos de diferentes escritas, idiomas e suportes documentais (Lose et al., 2024).

Entre as plataformas atualmente mais utilizadas destaca-se o Transkribus, desenvolvido inicialmente no âmbito do projeto europeu *tranScriptorium* e posteriormente integrado na *READ-COOP SCE* (Kahle et al., 2017; READ-COOP SCE, n/d). A plataforma combina tecnologias de HTR, análise de *layout* documental e gestão colaborativa de transcrições, permitindo o desenvolvimento de modelos adaptados

a diferentes tipos de documentação histórica. O sistema disponibiliza modelos públicos previamente treinados para múltiplos idiomas, mas possibilita igualmente a criação de modelos especializados a partir da construção de *ground truth*, entendido como o conjunto de imagens digitalizadas e transcrições alinhadas manualmente utilizado para o treino de modelos de reconhecimento automático (Muehlberger et al., 2019).

A utilização destas tecnologias tem ampliado significativamente o acesso a arquivos históricos, permitindo transformar documentação manuscrita em texto pesquisável e reutilizável em larga escala (Nockels et al., 2022), ao contrário da digitalização isolada que não torna os documentos pesquisáveis automaticamente, sendo necessário converter imagens em texto legível por máquina através de tecnologias de HTR (Prebor, 2024). Em contextos arquivísticos e patrimoniais, o HTR tem sido aplicado à transcrição de correspondência, registos administrativos, manuscritos religiosos e documentação científica, reduzindo o tempo necessário para leitura e indexação documental (Capurro et al., 2023; Muehlberger et al., 2019). Em Portugal, Silva (2025) destaca que a aplicação de IA em arquivos históricos pode contribuir para acelerar o acesso à informação e apoiar o trabalho arquivístico, embora estas abordagens permaneçam ainda pouco disseminadas em instituições portuguesas.

Contudo, a automatização da leitura paleográfica permanece condicionada pela qualidade do *corpus* documental e pela preparação dos dados de treino. Como observa Ciula (2017), a Paleografia Digital não corresponde apenas a uma substituição técnica da leitura realizada por especialista, mas uma contribuição para redefinir a metodologia na relação entre investigador, documento e representação textual. O treino de modelos HTR depende diretamente da intervenção humana na segmentação das linhas, transcrição paleográfica e validação dos resultados, exigindo processos contínuos de anotação e curadoria documental (Ackel, 2021; Spina, 2023). Como observa Caers (2024), os modelos de HTR dependem diretamente da qualidade das transcrições humanas utilizadas no treino, sendo a precisão paleográfica fundamental para o desempenho do reconhecimento automático. Assim, embora as ferramentas de IA ampliem significativamente as possibilidades de acesso e análise de manuscritos históricos, o reconhecimento automático continua dependente da mediação crítica do investigador, particularmente em documentação caracterizada por elevada heterogeneidade gráfica e estrutural.

Os arquivos hospitalares históricos constituem um conjunto documental particularmente relevante para os estudos sobre saúde, práticas médicas e administração institucional, reunindo informação clínica, administrativa e social produzida no contexto do internamento dos pacientes. Registos de admissão, papeletas médicas, livros de enfermaria e relatórios clínicos permitem reconstruir não apenas trajetórias individuais de doença, mas também práticas de observação médica, organização hospitalar e formas históricas de produção de informação sobre o corpo e a saúde (Risse & Warner, 1992; Hirst, 2018). Contudo, estes documentos apresentam frequentemente elevada complexidade paleográfica e estrutural, combinando diferentes grafias, abreviaturas técnicas, tabelas e anotações realizadas por múltiplos profissionais de saúde ao longo do internamento dos pacientes.

A utilização de tecnologias de HTR neste tipo de documentação permanece particularmente desafiante. Como demonstram Nockels et al., (2022), o desempenho dos modelos de reconhecimento automático varia significativamente conforme a regularidade gráfica, a organização do documento e a qualidade das imagens digitalizadas. Em registos hospitalares históricos, a coexistência de elementos pré-impressos e manuscritos, associada à diversidade caligráfica e terminológica, tende a dificultar tanto a segmentação automática do *layout* documental como o reconhecimento textual. Estas limitações tornam-se ainda mais

evidentes em documentação médica produzida no início do século XX, caracterizada por abreviaturas clínicas, irregularidades paleográficas e múltiplas intervenções manuscritas no mesmo documento.

Centrado nas papeletas médicas dos Hospitais da Universidade de Coimbra (HUC), preservadas no Arquivo da Universidade de Coimbra (AUC), o presente estudo, pretende analisar um *corpus* documental constituído por 174 registos hospitalares produzidos entre 1910 e 1936, correspondendo a aproximadamente 870 páginas digitalizadas referentes a indivíduos pertencentes à Coleção de Esqueletos Identificados (CEI) da Universidade de Coimbra (UC). Partindo deste *corpus*, o presente trabalho analisa a aplicação de modelos de HTR à documentação médico-administrativa portuguesa do início do século XX, discutindo as potencialidades e limitações da utilização de modelos generalistas disponíveis na plataforma Transkribus. Simultaneamente, procura refletir sobre o papel da mediação humana na construção de modelos especializados para arquivos hospitalares históricos, contribuindo para o debate contemporâneo sobre paleografia digital, Humanidades Digitais e acesso à informação em documentação manuscrita histórica.

Método

Com vista o cumprimento do objetivo proposta, numa uma primeira fase, realizou-se uma revisão da literatura especializada sobre digitalização de documentação histórica, HTR e aplicação de inteligência artificial em contextos arquivísticos e historiográficos. Posteriormente, procedeu-se à identificação, organização e digitalização dos documentos, de acordo com as orientações da *International Federation of Library Associations and Institutions* (IFLA) e do *International Council on Archives* (ICA) para digitalização de documentação histórica (McIlwaine et al., 2002). As imagens digitais foram posteriormente importadas para a plataforma Transkribus, desenvolvida para reconhecimento automático, transcrição e análise de manuscritos históricos através de tecnologias de HTR (Kahle et al., 2017; READ-COOP SCE, n/d).

Em 2025 foram realizados os testes preliminares utilizando os modelos públicos *Transkribus Portuguese Handwriting M2*, *Portuguese Handwriting 16th–19th century* e *Transkribus Print M1*. Para esta fase, selecionaram-se cinco papeletas médicas, correspondentes a aproximadamente 25 a 30 páginas digitalizadas. A preparação dos documentos incluiu a definição do *layout* das páginas, identificação das regiões textuais e segmentação das linhas de escrita. Procedeu-se igualmente à transcrição paleográfica manual das linhas selecionadas, respeitando a grafia original, abreviaturas e irregularidades presentes nos documentos históricos. Como observa Prebor (2024), os sistemas de HTR dependem diretamente da qualidade da segmentação e do alinhamento entre imagem e transcrição para o treino dos modelos.

No mês de maio de 2026, a pesquisa encontra-se em desenvolvimento a construção do *ground truth*, entendido como o conjunto de imagens digitalizadas e transcrições alinhadas manualmente utilizado para o treino de modelos de HTR adaptados a grafias específicas (Muehlberger et al., 2019). Este processo envolve a segmentação das linhas de texto e o respetivo alinhamento paleográfico das transcrições, permitindo a criação de um *corpus* de treino adaptado à documentação médico-administrativa portuguesa do início do século XX. As informações identificadas nos documentos incluem dados administrativos e clínicos, nomeadamente nome, idade, sexo, profissão, estado civil, morada, diagnósticos, tratamentos, tempo de internamento e causa de morte. O *corpus* inicial em

preparação é composto por aproximadamente 40 a 50 páginas transcritas manualmente, destinadas ao treino preliminar do modelo HTR.

Resultados

O *corpus* documental em análise é constituído por papeletas médicas produzidas pelos HUC, que correspondem a formulários médico-administrativos pré-impressos, posteriormente preenchidos manualmente por diferentes profissionais de saúde ao longo do internamento dos pacientes. Estes registos integram informação relativa 174 indivíduos, com idades compreendidas entre os oito e os 79 anos e períodos de internamento compreendidos entre um e 1131 dias. Estes indivíduos, cujos esqueletos estão hoje na CEI, tiveram entrada e faleceram nos HUC, permitindo a preservação da respetiva documentação clínica. A análise preliminar da documentação permitiu identificar elevada heterogeneidade estrutural e gráfica, verificando-se a coexistência de texto datilografado e manuscrito, diferentes organizações de página, estruturas tabulares e múltiplas grafias no mesmo documento.

Os testes preliminares realizados com modelos públicos de HTR treinados para documentação em língua portuguesa revelaram limitações significativas na transcrição automática deste tipo documental. Foram testados os modelos *Transkribus Portuguese Handwriting M2*, *Portuguese Handwriting 16th–19th century* e *Transkribus Print M1* em cinco papeletas médicas, correspondentes a aproximadamente 30 páginas digitalizadas. Os modelos testados e os respetivos resultados preliminares encontram-se sintetizados na Tabela 1.

Modelo	Tipo	Resultado preliminar
Transkribus Print M1	Texto impresso	Melhor desempenho em datilografado
Portuguese Handwriting 16th–19th	Manuscrito	Elevada taxa de erro
Portuguese Handwriting M2	Manuscrito	Dificuldades com múltiplas grafias

Tabela 1: Modelos de HTR testados na plataforma Transkribus e respetivo desempenho preliminar nas papeletas médicas dos HUC (Fonte: elaboração própria).

Verificaram-se dificuldades recorrentes na segmentação automática das linhas de texto (*layout*), sobretudo em páginas com elevada densidade de informação manuscrita ou com sobreposição entre texto manuscrito e elementos pré-impressos do formulário. O modelo *Transkribus Print M1* demonstrou melhor desempenho na leitura do texto datilografado, mas revelou incapacidade significativa na identificação e transcrição da escrita manuscrita. Conforme ilustrado na Figura 1, observaram-se erros frequentes na definição automática das linhas de texto, comprometendo subsequentemente o processo de reconhecimento textual.

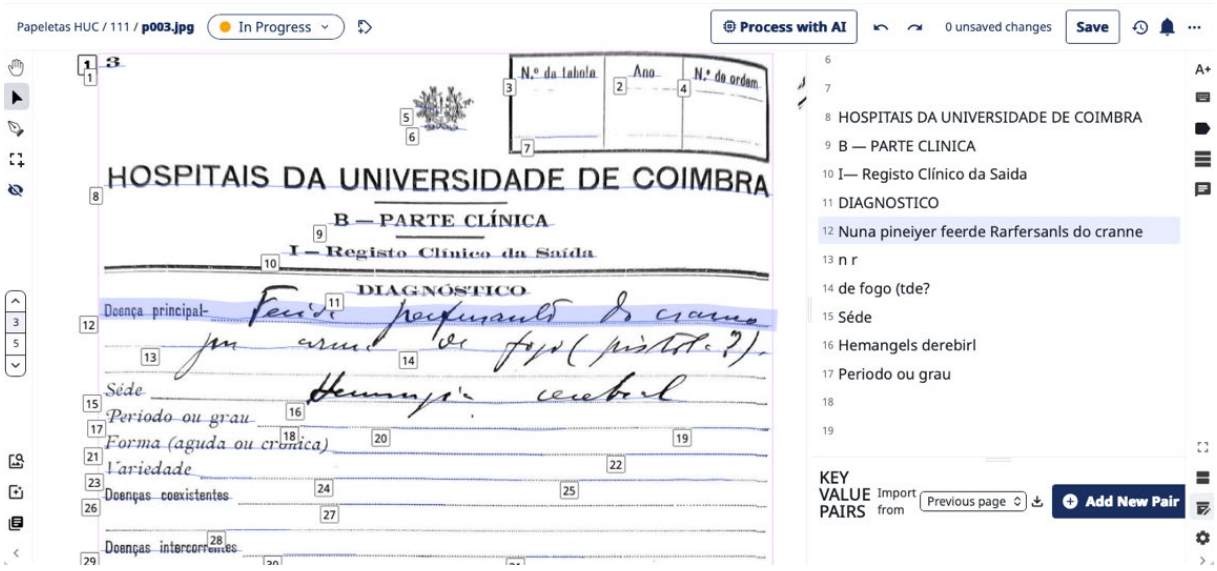

Figura 1: Exemplo de segmentação automática e transcrição realizada pelo modelo *Transkribus Print M1* em papeleta médica dos Hospitais da Universidade de Coimbra, evidenciando dificuldades na identificação das linhas de texto e no reconhecimento da escrita manuscrita. Em destaque a linha 12 com a transcrição realizada pela IA. (Formulário de registo, parte superior da página 3 – Registo Clínico de Saída. AUC, Fundo UC – HUC, Série: Papeletas de doentes (1870–1954), Vol. n.º 48/cx. 72, Depósito V – 2.ª D.).

Por sua vez, os modelos treinados para manuscritos em português apresentaram dificuldades na interpretação de abreviaturas médicas, grafias irregulares e terminologia clínica específica dos registos hospitalares do início do século XX. Verificou-se igualmente que um mesmo documento podia conter diferentes estilos caligráficos, resultantes do preenchimento por múltiplos profissionais de saúde ao longo do internamento do paciente, aumentando a variabilidade gráfica e dificultando o reconhecimento automático.

Embora não tenha sido realizada, nesta fase, uma avaliação quantitativa sistemática através de métricas como *Character Error Rate* (CER) ou *Word Error Rate* (WER), os resultados observados evidenciaram taxas de erro elevadas, particularmente em páginas caracterizadas pela coexistência de escrita manuscrita e datilografada, presença de tabelas e elevada densidade textual. Em diversos casos, a transcrição automática gerada pelos modelos revelou-se insuficiente para utilização direta, exigindo correção manual extensiva. Face às limitações observadas, a preparação do *corpus* de treino passou a privilegiar a criação manual do *layout* documental, permitindo maior controle sobre a segmentação das linhas de texto. Os processos automáticos foram reservados para páginas com estruturas mais regulares ou maior legibilidade gráfica.

Atualmente, o trabalho encontra-se na fase de construção do conjunto de dados de treino, obtido através da transcrição manual e alinhamento textual das papeletas médicas. Esta etapa envolve a definição das linhas de texto e transcrição das palavras manuscritas (Figura 2), respeitando a grafia original, abreviaturas e irregularidades paleográficas identificadas nos registos. O *corpus* inicial em preparação

é composto por aproximadamente 40 a 50 páginas transcritas, consideradas suficientes para o treino preliminar de um modelo específico adaptado a esta tipologia documental.

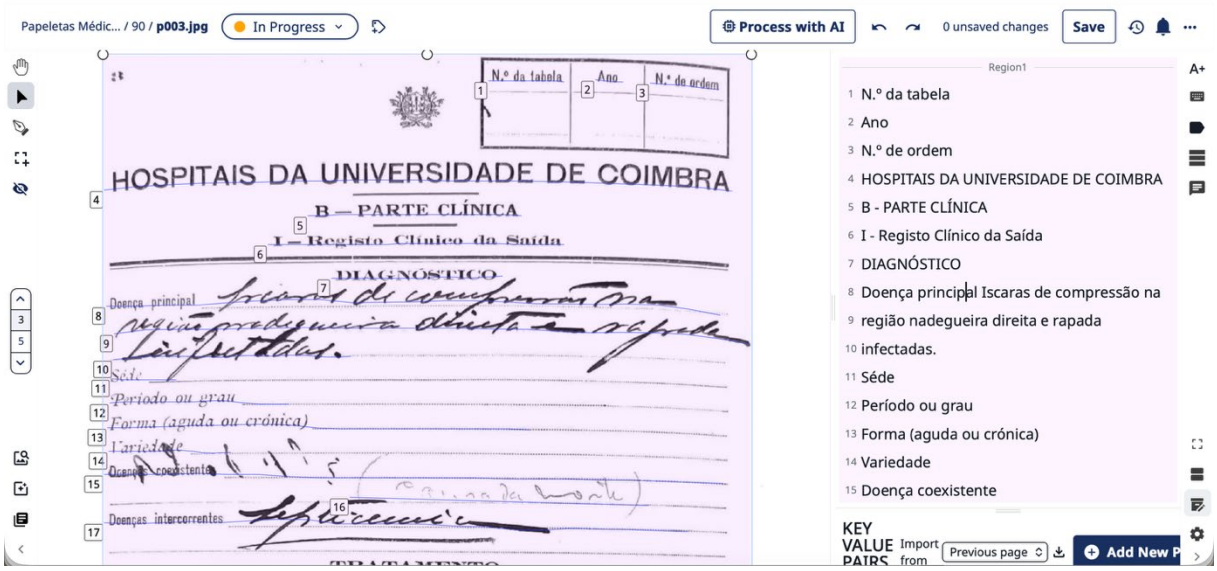

Figura 2: Processo de segmentação manual no Transkribus. A área destacada em rosa corresponde à região de texto adicionada (*Region 1*) e as linhas numeradas em azul representam o local das transcrições, ambas definidas manualmente pela autora para preparação do conjunto de dados de treino. A transcrição textual apresentada foi realizada pela equipa de investigação. (Formulário de registo, parte superior da página 3 – Registo Clínico de Saída. AUC, Fundo UC – HUC, Série: Papeletas de doentes (1870–1954), Vol. n.º 48/cx. 71, Depósito V – 2.ªD.)

Os resultados preliminares obtidos demonstram, assim, que a heterogeneidade estrutural e gráfica das papeletas médicas dos HUC constitui um obstáculo relevante à aplicação direta de modelos públicos de HTR treinados para documentação em língua portuguesa. Simultaneamente, evidenciam a necessidade de desenvolvimento de modelos especializados e da manutenção de forte intervenção humana durante as etapas de segmentação, transcrição e validação dos dados.

Discussão

A aplicação de tecnologias de HTR em documentos manuscritos tem sido apresentada como uma das transformações mais significativas das Humanidades Digitais e da Paleografia contemporânea, sobretudo pela possibilidade de ampliar o acesso, a pesquisa e a reutilização de documentação manuscrita em larga escala (Muehlberger et al., 2019; Nockels et al., 2022). Contudo, os resultados obtidos neste estudo demonstram que a automatização da leitura documental permanece condicionada pela materialidade específica dos manuscritos, pelas estruturas dos registos e pelas limitações dos modelos treinados.

Os problemas observados nas papeletas médicas dos HUC aproximam-se das dificuldades identificadas por Nockels et al. (2022) na revisão sistemática sobre a utilização do Transkribus em contextos patrimoniais. Os estudos demonstram que a aplicação de tecnologias de HTR depende fortemente das características paleográficas e estruturais de cada coleção documental, sobretudo em materiais compostos por múltiplas mãos, estruturas complexas e organização textual irregular. Neste contexto, os resultados obtidos não devem ser entendidos como uma falha isolada dos modelos utilizados, mas como consequência da própria heterogeneidade documental dos registos médico-administrativos históricos.

A coexistência de escrita datilografada e manuscrita evidencia igualmente uma limitação importante das abordagens generalistas de HTR. Embora plataformas como o Transkribus permitam integrar reconhecimento automático, análise de *layout* e treino de modelos personalizados (Kahle et al., 2017), a eficácia destes sistemas depende da estabilidade gráfica e estrutural dos documentos. Nas papeletas analisadas, a presença simultânea de formulários impressos, tabelas e anotações clínicas manuscritas produz um objeto documental híbrido, cuja complexidade ultrapassa os padrões normalmente utilizados no treino de modelos públicos. Prebor (2024), ao analisar a utilização do Transkribus em manuscritos hebraicos, demonstra que modelos genéricos tendem a apresentar taxas de erro significativamente elevadas quando aplicados a grafias e estruturas documentais distintas daquelas utilizadas no treino inicial. Apenas após a construção de modelos especializados adaptados ao *corpus* analisado foi possível reduzir substancialmente os valores de CER e WER. Embora desenvolvidos para outro contexto paleográfico, os resultados obtidos pela autora aproximam-se das dificuldades observadas nas papeletas médicas dos HUC, sobretudo no que se refere à necessidade de adaptação dos modelos às especificidades gráficas e estruturais da documentação histórica.

Os resultados dialogam igualmente com a reflexão de Ciula (2017), que defende que a Paleografia Digital não deve ser reduzida a uma mera automatização técnica da leitura, mas compreendida enquanto prática crítica e interpretativa. A autora ressalta que o ambiente digital transforma não apenas a velocidade de acesso aos manuscritos, mas também a própria relação entre investigador, documento e representação textual. Neste sentido, a necessidade de intervenção humana constante observada neste estudo reforça que o HTR não elimina a mediação paleográfica, antes, desloca-a para novas formas de interação entre leitura humana e processamento algorítmico.

Esta perspetiva aproxima-se igualmente das observações de Ackel (2021), para quem a Paleografia Digital representa uma convergência entre métodos tradicionais das ciências humanas e recursos computacionais. No presente estudo, essa realidade tornou-se particularmente evidente na construção manual do *ground truth*, etapa que exige decisões contínuas sobre grafias, abreviaturas, alinhamento textual e segmentação das linhas.

A dependência da intervenção humana confirma também a interpretação de Spina (2023), segundo a qual a introdução da IA nos fluxos de trabalho arquivísticos redefine o papel do especialista. Em vez de substituir o investigador, os sistemas de IA exigem acompanhamento permanente, revisão crítica e supervisão técnica. A automatização torna-se, assim, um processo colaborativo entre humano e máquina, no qual a qualidade dos resultados depende diretamente da curadoria documental e da preparação dos dados de treino. Caers (2024), ao discutir a utilização pedagógica do Transkribus, demonstra igualmente que os modelos de HTR permanecem diretamente dependentes da precisão das transcrições realizadas por técnicos especialistas utilizadas na construção dos modelos. O autor destaca que, embora a automatização reduza significativamente o tempo de transcrição, a qualidade do reconhecimento automático continua condicionada pela preparação paleográfica e pela anotação manual dos dados.

Do ponto de vista da Ciência da Informação e dos estudos arquivísticos, os resultados permitem igualmente discutir o papel das tecnologias digitais na ampliação do acesso aos arquivos históricos. Conforme argumenta Silva (2025) a aplicação de IA em manuscritos históricos apresenta potencial significativo para a recuperação da informação. Neste contexto, o presente estudo contribui para demonstrar não apenas o potencial destas tecnologias, mas também as exigências metodológicas necessárias para a sua implementação em documentação histórica portuguesa.

A utilização do Transkribus exigiu não apenas a digitalização das papeletas médicas, mas também um processo contínuo de segmentação textual, revisão paleográfica e validação manual das transcrições

produzidas. A perspetiva de "*life-cycle data curation*" defendida por Marsh et al., (2019) permite compreender que a digitalização documental constitui apenas uma etapa dentro de um processo mais amplo de gestão, preservação, mediação e acesso à informação. Nesse contexto, a utilização do Transkribus não se limita ao reconhecimento automático da escrita, envolvendo igualmente organização e preparação da documentação, revisão paleográfica e curadoria contínua dos dados produzidos. A heterogeneidade do *corpus* — composto por formulários dactilografados preenchidos com anotações manuscritas de diferentes profissionais — exigiu uma supervisão humana constante ao longo de todas as etapas de tratamento documental.

A inexistência, até ao momento, de métricas quantitativas formais como CER ou WER impede comparações diretas com estudos desenvolvidos noutros contextos documentais. Ainda assim, os problemas identificados durante os testes preliminares aproximam-se das dificuldades descritas por Capurro et al., (2023), sobretudo no que se refere à heterogeneidade gráfica dos manuscritos e à necessidade de adaptação dos modelos ao *corpus* documental específico. Mais do que um obstáculo técnico, estas limitações evidenciam uma questão central das Humanidades Digitais contemporâneas: a automatização da leitura histórica depende da construção prévia de *corpora* cuidadosamente preparados, anotados e interpretados por especialistas. Assim, os resultados obtidos reforçam que a IA aplicada à Paleografia Digital não substitui o conhecimento humano, mas exige novas formas de mediação crítica entre documento, tecnologia e interpretação histórica.

Conclusões

Os resultados preliminares deste estudo demonstram que a aplicação de modelos generalistas de HTR às papeletas médicas dos HUC apresenta limitações significativas, sobretudo devido à heterogeneidade gráfica e estrutural dos documentos analisados. A coexistência de escrita datilografada e manuscrita, associada à multiplicidade de mãos e à organização tabular dos registos, confirmou a necessidade de desenvolvimento de um modelo específico para documentação médica portuguesa do início do século XX.

O estudo permitiu igualmente verificar que a automatização da leitura documental depende diretamente da preparação e curadoria humanas dos dados de treino. A criação manual do *layout*, a segmentação das linhas e a transcrição paleográfica continuam a desempenhar um papel central no processo de construção do *corpus*, demonstrando que a IA não substitui a leitura interpretativa, mas funciona como ferramenta complementar de apoio ao acesso e tratamento da informação histórica.

Neste sentido, a investigação confirma o potencial das Humanidades Digitais e da Paleografia Digital para ampliar o acesso a arquivos manuscritos, mas evidencia igualmente a necessidade de modelos treinados especificamente para conjuntos documentais complexos. A continuidade do trabalho passará pela expansão do *ground truth* e pelo treino de um modelo HTR adaptado às papeletas médicas dos HUC, permitindo posteriormente avaliar quantitativamente o seu desempenho através de métricas de erro e precisão.

A longo prazo, a criação de modelos especializados para documentação clínica histórica poderá contribuir para a preservação ativa do património documental e para a transformação de arquivos

manuscritos em informação pesquisável e reutilizável, reforçando o papel dos arquivos como espaços de produção, mediação e difusão do conhecimento científico.

Referências bibliográficas

Ackel, A. (2021). Abordagens digitais para estudos de paleografia de paleografia: Desafios, atualidade, desdobramentos. *LaborHistórico*. https://orcid.org/0000-0002-8283-4417

Barros, J. D. (2019). Fontes históricas: Introdução aos seus usos historiográficos. *Anais Do 2° Encontro Internacional da ANPUH (História & Parcerias)*. https://www.historiaeparcerias2019.rj.anpuh.org/resources/anais/11/hep2019/1569693608_ARQUIVO_bd3da9a036a806b478945059af9aa52e.pdf

Borges, L. C., & Silva, A. D. (2019). Transcrições em linha: E-learning de Paleografia em arquivos europeus. *Revista Portuguesa de História*, *49*, 31–54. https://doi.org/10.14195/0870-4147_49_2

Borges, L. C., Silva, A. D., & Espírito Santo, S. M. (2023). Arquivos como fontes de poder, marginalização e silêncios em Portugal e no Brasil. In *Atas do 14.º Congresso Nacional de Bibliotecários, Arquivistas e Documentalistas* [Faro, Portugal]. https://doi.org/10.48798/congressobad.2984

Caers, B. (2024). Teaching handwritten text recognition: Can new technologies save old skills? *Quaerendo 54(2-3), 198-209.* https://doi.org/10.1163/15700690-bja10024

Ciula, A. (2017). Digital palaeography: What is digital about it? *Digital Scholarship in the Humanities*, 32(2), ii89–ii105. https://doi.org/10.1093/llc/fqx042

Capurro, C., Provatorova, V., & Kanoulas, E. (2023). Experimenting with training a neural network in Transkribus to recognise text in a multilingual and multi-authored manuscript collection. *Heritage*, *6*(12), 7482-7494. https://doi.org/10.3390/heritage6120392

Cornelius, I. (2002). Theorizing information for information science. In *Annual Review of Information Science and Technology, 36*(1). https://doi.org/10.1002/aris.1440360110

Dacos, M. (2011). Manifesto das humanidades digitais. *Humanidades Digitais*: Grupo de Pesquisas | Universidade de São Paulo. https://humanidadesdigitais.org/manifesto-das-humanidades-digitais/

Hirst, C. (2018). HHARP: The historical hospital admission records project – a review. *Internet Archaeology*, *47*. https://doi.org/10.11141/ia.47.6

McIlwaine, J., Comment, J. M., Wolf, C. D., Peters, D., Justrell, B., Varlamoff, M. T., & Koopman, S. (2002). *Guidelines for digitalization projects for collections and holdings in the public domain, particularly those held by libraries and archives*. IFLA. https://www.ifla.org/wp-content/uploads/2019/05/assets/preservation-and-conservation/publications/digitization-projects-guidelines.pdf

Kahle, P., Colutto, S., Hackl, G., & Mühlberger, G. (2017). Transkribus: A service platform for transcription, recognition and retrieval of historical documents. In *14th IAPR International Conference on Document Analysis and Recognition*. pp. 19-24. https://doi.org/10.1109/ICDAR.2017.307

Kirschenbaum, M. G. (2010). What is digital humanities and what's it doing in English departments? *ADE Bulletin*, 55–61. https://doi.org/10.1632/ade.150.55

Le Coadic, Y.-F. (1996). *A ciência da informação*. Briquet de Lemos.

Le Goff, J. (1990). *História e memória*. Editora da Unicamp.

Lose, A. D., Santos, J. G., Jesus, L. C., Magalhães, L., & Xavier, L. (2024). Transkribus: Uma ferramenta de paleografia digital mediando pesquisas em fontes inquisitoriais. *LaborHistórico*, 10(1). https://doi.org/10.24206/lh.v10i1.63285

Marsh, D. E., St. Andre, S., Wagner, T., & Bell, J. A. (2023). Attitudes and uses of archival materials among science-based anthropologists. *Archival Science*. https://doi.org/10.1007/s10502-023-09411-z

Marsh, D. E., Punzalan, R. L., & Johnston, J. A. (2019). Preserving anthropology's digital record: CoPAR in the age of electronic fieldnotes, data curation, and community sovereignty. *The American Archivist*, 82(2), 268-302. https://doi.org/10.17723/aarc-82-02-01

Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., ... & Zagoris, K. (2019). Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, 75(5), 954-976.

Nockels, J., Gooding, P., Ames, S., & Terras, M. (2022). Understanding the application of handwritten text recognition technology in heritage contexts: A systematic review of Transkribus in published research. *Archival Science*, *22*(3), 367–392. https://doi.org/10.1007/s10502-022-09397-0

Parezo, N. (1999). Preserving anthropology's heritage: CoPAR, anthropological records, and the archival community. *The American Archivist*, *62*(2), 271–306. https://www.jstor.org/stable/40294124

Prebor, G. (2024). From digitization and images to text and content: Transkribus as a case study. *Proceedings of the Association for Information Science and Technology,* 60, 1102-1103. https://doi.org/10.1002/pra2.958

READ-COOP SCE. (n/d). Transkribus: Unlocking the past with AI. Transkribus. https://www.transkribus.org

Risse, G. B., & Warner, J. H. (1992). Reconstructing clinical activities: Patient records in medical history. *Social History of Medicine*, *5*(2), 183–205. https://doi.org/10.1093/shm/5.2.183

Sanjad, N. (2017). Exposições internacionais: Uma abordagem historiográfica a partir da América Latina. *História, Ciências, Saúde-Manguinhos*, 24(3), 785-826. https://doi.org/10.1590/S0104-59702017000300013

Silva, A. D. (2025). A Inteligência Artificial no acesso à informação em documentos manuscritos. *Cadernos BAD*, 1-2. https://doi.org/10.48798/cadernosbad.3095

Spina, S. (2023). Artificial Intelligence in archival and historical scholarship workflow: HTS and ChatGPT. *arXiv preprint*. https://doi.org/10.48550/arXiv.2308.02044

Zeitlyn, D. (2012). Anthropology in and of the archives: Possible futures and contingent pasts – Archives as anthropological surrogates. *Annual Review of Anthropology*, *41*(1), 461–480. https://doi.org/10.1146/annurev-anthro-092611-145721