EDITORIAL

Cadernos BAD, 2025, 1-2
https://doi.org/10.48798/cadernosbad.3095

A Inteligência Artificial no acesso à informação em documentos manuscritos[1]

Ana Margarida Dias da Silva
Universidade de Coimbra, Centro de História da Sociedade e da Cultura

Paleografia digital

Acesso à informação

Comunicação da informação

Handwritten Text Recognition

Arquivos

A prática da paleografia de leitura tem acompanhado a evolução das novas tecnologias, atualmente também fazendo uso da Inteligência Artificial (IA). A Paleografia Digital é um campo fértil para auxiliar os arquivos e arquivistas na captação de informação, nomeadamente através do Reconhecimento de Texto Manuscrito (Handwritten Text Recognition - HTR), mas tem pouca ou nenhuma utilização em Portugal. O trabalho em apreço apresenta-se como um estudo exploratório que procura, na relação entre as funções de comunicação e acesso à informação com a Paleografia Digital, um método de trabalho em instituições detentoras de manuscritos, que acelere o acesso à informação, nomeadamente com a utilização do modelo de transcrição automática "Portuguese Handwriting 16th-19th century", desenvolvido na plataforma Transkribus. Este modelo foi criado por uma equipa de paleógrafos portugueses e brasileiros liderados por Hervé Baudry no âmbito do projeto Transcrever os Processos da Inquisição Portuguesa (1536-1821) (TraPrInq), projeto exploratório subsidiado pela FCT (ref. HAR-HIS/0499/2021), que decorreu de 17.1.2022 a 16.7.2023 no CHAM-Centro de Humanidades.

Artificial Intelligence for accessing information in handwritten documents

Digital palaeography

Access to information

Communication of information

Handwritten Text Recognition

Archives

The practice of reading palaeography has kept pace with the evolution of new technologies, currently also making use of Artificial Intelligence (AI). Digital palaeography is a fertile field for helping archives and archivists capture information, particularly through Handwritten Text Recognition (HTR), but it has little or no use in Portugal. This work is presented as an exploratory study that seeks, in the relationship between the functions of communication and access to information with Digital Palaeography, a method of working in institutions holding manuscripts that speeds up access to information, namely by using the automatic transcription model 'Portuguese Handwriting 16th-19th century', developed on the Transkribus platform. This model was created by a team of Portuguese and Brazilian palaeographers led by Hervé Baudry as part of the project Transcribing the Processes of the Portuguese Inquisition (1536-1821) (TraPrInq), an exploratory project subsidised by FCT (ref. HAR-HIS/0499/2021), which ran from 17.1.2022 to 16.7.2023 at the CHAM-Humanities Centre.

[1] Este trabalho é financiado por fundos nacionais através da FCT – Fundação para a Ciência e a Tecnologia, I.P., no âmbito do projeto UIDB/00311/2020 com o identificador DOI 10.54499/UIDB/00311/2020 DOI https://doi.org/10.54499/UIDB/00311/2020

INTRODUÇÃO

Desde sempre as administrações sentiram a necessidade de ler e transcrever documentos antigos o que levou à especialização de profissionais numa técnica designada por "paleografia de leitura", que ao longo dos séculos se manteve de forma consistente, fazendo jus à etimologia do conceito (BORGES & SILVA, 2018, p. 3).

De paleografia empírica, no sentido restrito de prática de decifrar escritas/documentos antigos (NUNES, 1973, p. 9), a ciência auxiliar, primeiro da Diplomática, depois da História, é só no século XX, que a paleografia (pese embora a etimologia da palavra) vê renovados o seu objeto de estudo e método, atingindo assim a estatuto de ciência autónoma (BORGES & SILVA, 2018, p. 35).

O século XXI trouxe novos usos e uma nova vida à ciência paleográfica, aproximando-a de um público mais vasto, sabendo aproveitar as potencialidades da *Internet* e da *Web 2.0*. Vejam-se, por exemplo, os inúmeros cursos de *e-learning* de ensino e prática da paleografia de leitura em arquivos europeus (BORGES & SILVA, 2018) ou o crescente número de projetos de *crowdsourcing* e participação cidadã de transcrição massiva de documentos, que permitem que qualquer pessoa com acesso à Internet participe na transcrição e leitura de manuscritos (SILVA & BORGES, 2018).

Na verdade, a prática da paleografia de leitura tem acompanhado a evolução das novas tecnologias, atualmente também fazendo uso da Inteligência Artificial (IA). A Paleografia Digital é um campo fértil para auxiliar os arquivos e arquivistas na captação de informação, nomeadamente através do Reconhecimento de Texto Manuscrito (HTR), mas tem pouca ou nenhuma utilização em Portugal.

A maior parte da documentação custodiada em arquivos é descrita ao nível da série, pois o grau de exaustividade e especificidade necessários para a descrição ao nível da peça não se coadunam nem com as capacidades humanas nem com as possibilidades financeiras das instituições, pela morosidade que isso implicaria. Isso provoca uma quantidade significativa de massa documental não tratada e é neste ponto que a Inteligência Artificial pode ser uma ferramenta útil aos arquivos e aos arquivistas.

O desenvolvimento de programas de reconhecimento óptico de carateres (OCR) para material impresso foi momento marcante nas Humanidades Digitais, hoje de utilização fácil e corrente. Desde os anos 1990, com base em imagens digitais e IA, foram iniciados projetos de reconhecimento óptico de manuscritos com o objetivo de remover "the manuscript from the matrix of its paper model" (STUTZMANN, 2011, p. 219). Em 2013, a mesma cientista anunciou o triunfo da paleografia digital, apesar da complexidade da tarefa e do muito que ainda estava por conseguir (STUTZMANN, 2017).

A plataforma Transkribus foi criada por investigadores da universidade de Innsbruck, na Áustria, e utiliza o HTR em conjunto com recursos de inteligência artificial para reconhecer "a morfologia das letras, as características linguísticas de cada língua, bem como o modus scribendi das mãos ali representadas e faça, com celeridade, transcrições automatizadas." (Lose et al. 2024, 9).

O trabalho em apreço apresenta-se como um estudo exploratório que procura, na relação entre as funções de comunicação e acesso aos documentos e à informação com a Paleografia Digital, um método de trabalho nas instituições detentoras de manuscritos que acelere o acesso à informação, nomeadamente com a utilização do modelo de transcrição automática "Portuguese Handwriting 16th-19th century".

1. O modelo de transcrição automática "Portuguese Handwriting 16th-19th century"

O «TraPrInq Portuguese Handwriting 16th-19th c.» é um modelo genérico criado no âmbito do Projeto TraPrInq (01.2022 a 07.2023), financiado pela FCT, por uma equipa luso-brasileira de paleógrafos: Hervé Baudry, Susana Tavares Pedro, Carla Vieira, Jorge Ferreira Paulo, Leonor Dias Garcia, Ana Margarida Dias da Silva, Maria Olinda Alves Pereira, Mário Soares Fatela, Marize Helena de Campos, Natalia Casagrande Salvador, Suzana Maria de Sousa Santos Severs. Este modelo HTR, disponibilizado em acesso aberto, baseou-se nos registos dos julgamentos da Inquisição Portuguesa produzidos entre 1536 (alguns documentos ainda antes) e 1821, ou seja, são 250 épocas de treino. Contém uma transcrição cuidadosa de 6.226 páginas (Validation Set (VS): 505 pp.; Training Set (TS): 5.721 pp.) extraídas de 830 processos, principalmente do Tribunal de Lisboa, num total de 1.268.040 palavras (VS: 107.760 palavras; TS: 1.160.280). Significa que o CER do VS set é de 5,2%. A transcrição reproduz a ortografia das palavras e das abreviaturas, utiliza caracteres especiais para os sinais de abreviatura de base e um único MACRON DE COMBINAÇÃO para todos os sinais de abreviatura sobrescritos, e moderniza a separação das palavras.

2. Transcrição automática de manuscritos portugueses

A primeira nota reter é a de que, apesar do modelo "Portuguese Handwriting 16th-19th century" ter sido criado a partir dos processos da Inquisição Portuguesa, e o conteúdo poder ser diferente de outros manuscritos, há constantes que facilitam a utilização do modelo: estarem datados do mesmo período temporal (séculos XVI a XIX), conterem nomes portugueses idênticos (quer nomes próprios, quer apelidos) e localidades e geografias de Portugal.

Numa experiência utilizando imagens de registos paroquiais portugueses no Transkribus, com a aplicação do modelo "Portuguese Handwriting 16th-19th century", tornado público em setembro de 2023 (Silva, 2025), verifica-se uma boa leitura da grafia do século XVI (fig. 1) e dos séculos XVII (fig. 2) e XVIII (fig. 3).

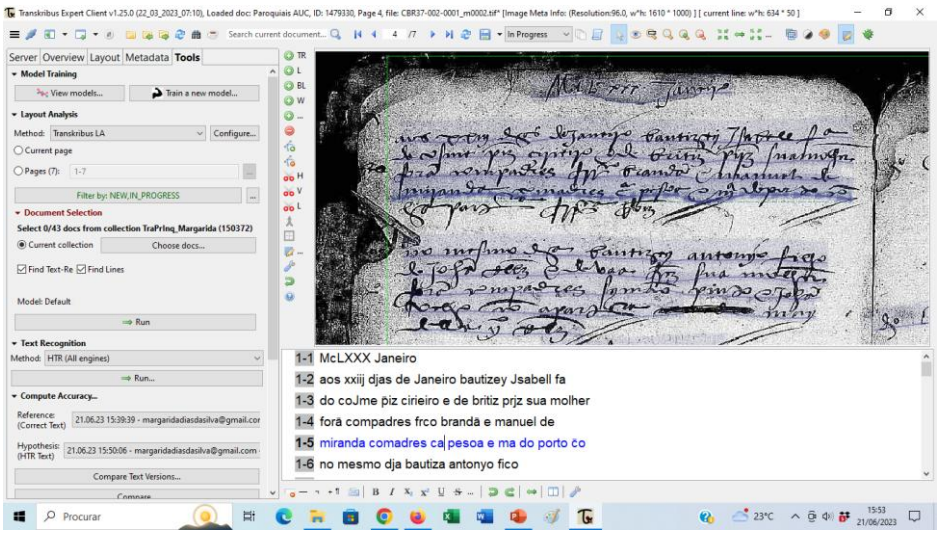

Figura 1 - Livro de Batismos, paróquia de São Tiago, Coimbra, 1530 (CBR37-002-0001_m0002), Arquivo da Universidade de Coimbra

Outra grande vantagem do Transkribus, é a possibilidade de introdução de metadados e de colocar etiquetas (*tags*), como Pessoa, Local ou Data, por exemplo, e, ainda, de acrescentar propriedades (no caso da Pessoa, a Idade, o Pseudónimo, entre outros) (fig. 2 e 3).

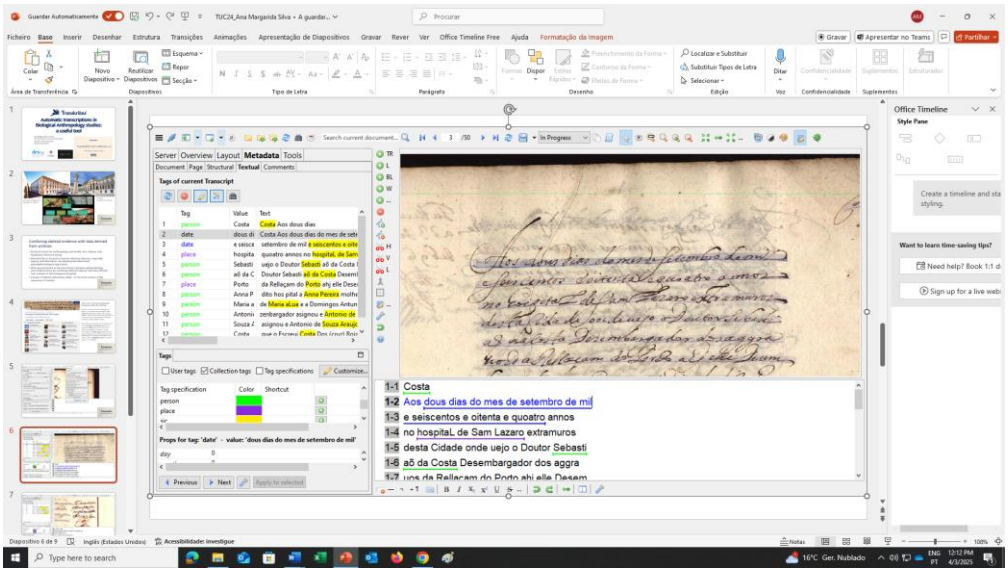

Figura 2 - Admissão de doentes no Hospital de S. Lázaro (1684), Arquivo da Universidade de Coimbra

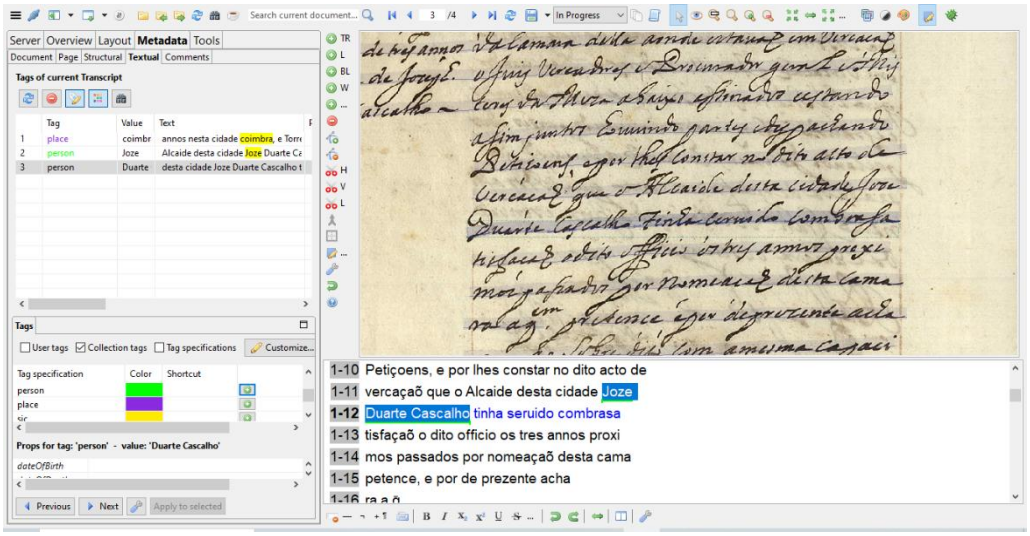

Figura 3 - Atas das vereações (1702), Arquivo Municipal de Coimbra

Os dados extraídos são facilmente exportados para Word ou PDF que depois podem ser, por sua vez, exportados para as plataformas de descrição arquivística on-line.

Considerações finais

A democratização do acesso à informação tem sido conseguida através da disponibilização de conteúdos on-line, acessíveis 24 horas por dia, todos os dias, a quem tenha acesso a um computador com Internet. A crescente utilização de plataformas colaborativas da Web 2.0 ou de IA para a transcrição massiva de documentos têm dado um novo uso à paleografia, cujo conhecimento é essencial para a leitura de manuscritos. A participação e colaboração de todos permitem um maior grau de exaustividade da descrição arquivística, que de outra forma não seria possível nos arquivos, tornando a participação fundamental na recuperação da informação.

Nos exemplos acima citados, verifica-se uma maior velocidade e rapidez na transcrição e a

possibilidade de indexar e pesquisar nomes, locais, datas, profissões, entre outros, abre boas perspetivas para pesquisas automática.

No entanto, a transcrição automática de manuscritos também levanta questões éticas: quem transcreveu o documento? A quem é dada a autoria da transcrição? Será que a Paleografia Digital transforma os paleógrafos em meros corretores de dados? Ou, pelo contrário, o paleógrafo sai valorizado pelos conhecimentos de leitura de manuscritos?

O processo de transcrição automática acelera o acesso à informação, embora não descure a existência de profissionais com conhecimentos de leitura paleográfica, sobretudo para corrigir a máquina. Em suma, a IA não dispensa o elemento humano.

Designação do projeto/iniciativa

Projeto Transcrever os Processos da Inquisição Portuguesa (1536-1821) (TraPrInq), projeto exploratório subsidiado pela FCT (ref. HAR-HIS/0499/2021), que decorreu de 17.1.2022 a 16.7.2023 no CHAM-Centro de Humanidades. Para mais informações, consultar a página do projeto: https://traprinq.mozellosite.com/

Público-alvo

Esta apresentação destina-se a todos os profissionais de informação que trabalham com documentos manuscritos.

Ligações web úteis

Ligação web ao modelo "Portuguese Handwriting 16th-19th century" no Transkribus: https://readcoop.eu/model/portuguese-handwriting-16th-19th-century/

Ligação web para os dados: https://traprinq.mozellosite.com/dados/

Referências bibliográficas

Borges, L.C., & Silva, A.M.D. da (2018). Transcrições em linha: e-learning de Paleografia em arquivos europeus. *Revista Portuguesa de História*, XLVIII, 39-59. https://doi.org/10.14195/0870-4147_49_2

Lose, A.D., Santos, J.G.V.A. dos, Jesus, L.C.M. de, Magalhães, L.B.S., & Lucia Furquim Werneck Xavier, L.F.W. (2024). Transkribus: uma ferramenta de paleografia digital mediando pesquisas em fontes inquisitoriais. *Revista LaborHistórico*, 10(1): e63285. https://doi.org/10.24206/lh.v10i1.63285

Nunes, E.B. (1973). O conceito novo de paleografia. *Portugaliae Historica*, I, 9-12.

SILVA, A.M.D. (2025). Transcrições automáticas nos arquivos distritais portugueses: acelerar o acesso à informação. *Cultura. Revista de História e Teoria das Ideias*, 41-42 (2023), 269-286.

STUTZMANN, E. (2011). Nouvelles technologies au service de la codicologie et de la paléographie. *Scriptorium*, 65(1), 217-223.

STUTZMANN, E. (2017). Paléographie, une révolution numérique. *L'Histoire*, 439 (septembre).

Silva, Ana Margarida Dias da (2025). "A Inteligência Artificial no acesso à informação dos documentos manuscritos". Cadernos BAD, n. 1-2. https://doi.org/10.48798/cadernosbad.3095

Conflitos de Interesse
A autora declara a inexistência de conflitos de interesse na realização do presente trabalho.

Revisão por Pares
Esta revista usa um sistema de revisão duplamente cega por pares assegurada pelo conselho científico da Cadernos BAD.

Confidencialidade dos Dados
A autora declara ter seguido os protocolos de RGPD.

Financiamento, apoio e patrocínios
Este trabalho é financiado por fundos nacionais através da FCT – Fundação para a Ciência e a Tecnologia, I.P., no âmbito do projeto UIDB/00311/2020 com o identificador DOI 10.54499/UIDB/00311/2020 DOI https://doi.org/10.54499/UIDB/00311/2020

CENTRO DE HISTÓRIA DA SOCIEDADE E DA CULTURA

FCT Fundação para a Ciência e a Tecnologia

Recebido	Aceite	Publicado
03/04/2025	17/04/2025	27/07/2025