Dossiê | Humanidades Digitais e Documentos Históricos: Transcrever, Catalogar, Editar

Apresentação

Humanidades Digitais e Documentos Históricos: Transcrever, Catalogar, Editar

Hervé Baudry, Susana Tavares Pedro***

Cultura. Revista de História e Teoria das Ideias 41-42 (2023), 235-246. ISSN 0870-4546.
URL: https://revistas.fcsh.unl.pt/cultura/article/view/1223

O reconhecimento de texto

Os artigos que constituem o dossier intitulado "As Humanidades Digitais[1] e Documentos Históricos: Transcrever, Catalogar, Editar" referem frequentemente a tecnologia do reconhecimento textual automatizado. Com predominância da língua inglesa nas aplicações de fácil acesso usadas na Internet, a expressão mais generalizada relativa aos procedimentos é a *Optical Character Recognition* (OCR), reconhecimento óptico de carateres (Wang 2024; Wang 2021)[2]. Trata-se de tornar a imagem de um texto impresso num texto eletrónico com extensão doc, odt ou rtf. Entendeu-se assim que *character* reenvia só para os carateres de imprensa, o que, linguisticamente, é redutor. Após um período de investigações avançadas (Stutzmann 2017, 30-31; Santos Ruiz 2017; Tarte 2014, 112-135; Ciula 2005), há menos de dez anos apareceram novas aplicações capazes de realizar o mesmo tipo de operação com imagens de textos manuscritos, ou *Handwritten Text Recognition* (HTR). No entanto, a sigla OCR é utilizada por alguns no sentido do HTR, assumindo o reconhecimento dos caracteres escritos e impressos. Foi proposta a expressão *Automatic Text Recognition* (ATR) que unifica o reconhecimento automático dos manuscritos e dos impressos (Romein 2020). Fica esta a tendência geral: OCR e HTR referem o reconhecimento de texto, no primeiro caso, impresso, e no segundo, manuscrito.

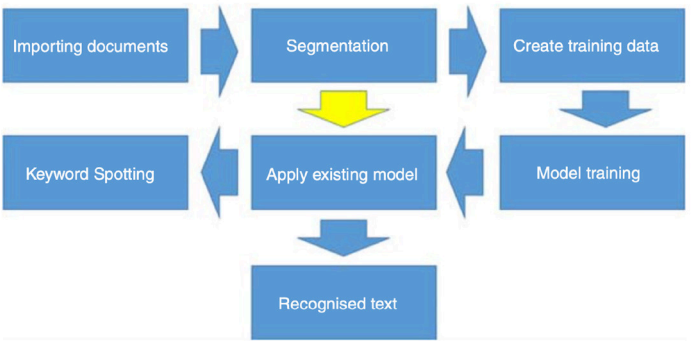
* CHAM, Faculdade de Ciências Sociais e Humanas, FCSH, Universidade NOVA de Lisboa, 1069-061 Lisboa, Portugal.
ORCID iD: https://orcid.org/0000-0001-9102-913X. *E-mail*: hbaudry@fcsh.unl.pt.
** CH-ULisboa – Centro de História Universidade de Lisboa, Portugal.
ORCID iD: https://orcid.org/0000-0003-2724-3187. *E-mail*: susana.t.pedro@gmail.com

1 Definidas por Dan Cohen como "uma comunidade de práticas dentro do formalismo académico: dos exemplos internacionais ao caso português" (Alves 2016, 91).

2 Sobre o desempenho comparado de cinco plataformas no reconhecimento de impressos antigos (ABBYY, eScriptorium, Nanonets, Tesseract, Transkribus), ver Sefil 2024.

Vários artigos deste número resultam de trabalhos realizados na plataforma de paleografia digital Transkribus[3] (Nockels 2022, 367-392; Mühlberger 2019, 954-976). Trata-se de uma aplicação em linha (*Software as a service*, SaaS) gerida pela sociedade cooperativa de direito europeu ReadCoop SCE[4]. Baseada na Universidade de Innsbruck (Áustria), foi fundada em 2019 e contava, nos finais de 2024, mais de 300 000 utilizadores particulares e institucionais, como arquivos, bibliotecas e universidades[5]. Transkribus providencia os meios que permitem a transcrição e a extração de informação[6] em massa de documentos manuscritos e impressos. Ao contrário dos programas não-SaaS, que podem ser instalados num computador e funcionam fora de conexão (*offline*)[7], Transkribus não permite o acesso aos algoritmos. Tudo o resto, em particular os dados, fica privado ou é tornado público conforme a vontade do utilizador.

É necessário descrever de maneira sucinta o *workflow* habitual no uso da plataforma após a abertura de uma conta. Como mostra o esquema seguinte (**Fig.1**), o material, constituído por imagens de resolução e qualidade suficientes[8], é importado em coleções e submetido ao motor PyLaia através de um modelo de HTR treinado com a Inteligência Artificial (rede neuronal e algoritmos) e a aprendizagem automática (*machine learning*).



1 Cf. Mühlberger 2019, 954.

[3] URL: https://www.transkribus.org/.

[4] "A cooperative to unlock our written past". https://readcoop.org/

[5] Entre os numerosíssimos projetos, limitemo-nos a indicar dois (com impressos): Bazzaco 2022; Couture 2022.

[6] Ver, por exemplo, Lang 2025; Ehrmann 2021.

[7] Por exemplo, eScriptorium (Kiessling 2019).

[8] Parte do trabalho preparativo pode ser relativo ao tratamento da imagem, aspeto que deixamos de lado nesta apresentação. É uma questão fundamental, bem antes do HTR ter sido tornado público (Roling 2020; Zheng 2004).

Já existem milhares de modelos de reconhecimento de texto na plataforma Transkribus, mas, até ao presente, só menos de três centenas foram tornados públicos pelos seus autores. Cada um tem um raio de ação mais ou menos largo. A maioria dos modelos foi treinada (*model training*, i.e. "criados") para transcrever textos redigidos com uma determinada caligrafia (uma "mão"), como os manuscritos de Jeremy Bentham ou de Michel Foucault. Qualquer pessoa que escreve corresponde a uma mão. Outros, chamados de modelos genéricos (Rabus 2019), foram treinados para transcrever um leque mais ou menos vasto de mãos, conforme o número de redatores, o período de produção dos documentos, as línguas, etc. Aos modelos de reconhecimento de texto devem juntar-se tipos de modelos específicos, em particular os modelos de segmentação (*layout*). São treinados para reconhecer as estruturas complexas, como, para dar só alguns exemplos, os textos manuscritos com múltiplas formas de *marginalia*, orientações várias, ou, no caso dos impressos, os periódicos com várias colunas de texto, ou ainda, nos documentos administrativos, as tabelas das fichas pré-impressas.

Quando o objetivo do utilizador é obter um texto transcrito, a descrição em dois passos, da importação à obtenção do texto transcrito, é o ideal, o que não significa o impossível. De precisar que, ao segundo passo, se segue a exportação dos dados, em particular em formato xml. Aqui, o ideal supõe o uso do modelo adequado — o que se produz cada vez mais com a extensão do número de modelos genéricos e o desenvolvimento dos super-modelos com os *Large Language Models*, LLM (Humphries, 2024) — com textos com *layout* simples ou já treinado. O certo é que há cada vez melhores surpresas[9], mas sim por utilizadores particulares, equipas de projetos de investigação ou de instituições públicas ou privadas. Neste caso, o *workflow* segue diversas etapas, as principais sendo as seguintes: a importação (até 5000 pdf), a segmentação (das páginas para HTR ou para treinar um modelo de *layout*), a transcrição manual dos textos (para produzir texto correto, ou *Ground Truth*, GT), os metadados, em particular a etiquetagem (*tagging*). Com uma quantidade de dados relevante, pode-se treinar o modelo, uma operação que terá de ser repetida, acrescentando novas transcrições GT conforme a amplitude do modelo desejado. A certa altura, no caso de modelos genéricos ambiciosos, a fase em duas etapas torna-se rotineira como parte do *workflow* quando a transcrição manual para produzir GT é facilitada, ou até substituída, pela transcrição automática, de melhor ou pior qualidade, conforme a evolução do próprio modelo nas fases sucessivas de treino.

Considera-se que um modelo é robusto, utilizável com ganhos de tempo e de legibilidade, quando transcreve textos com uma elevada exatidão. Neste caso, o indicador é

[9] Exceto o Transformer e alguns grandes modelos, geralmente multilingues, treinados com modelos públicos.

a taxa média de erro (*Character Error Rate*, CER[10]). Um CER abaixo de 10% (ou seja, uma taxa de exatidão de 90%) é um bom indicador, sendo procuradas taxas abaixo de 5%. Costuma-se dizer que o CER de 0% não é atingível. Só o futuro o confirmará ou não.

O projeto "Transcrever os processos da Inquisição Portuguesa, 1536-1821"

Um modelo treinado de HTR tem duplo uso: transcrever textos e servir de modelo base para modelos mais performativos na plataforma. Foi este propósito que moveu a equipa do projeto "Transcrever os processos da Inquisição Portuguesa, 1536-1821 (TraPrInq)"[11]. Durante 18 meses, onze historiadores paleógrafos[12] transcreveram mais de 6000 páginas de processos digitalizados da Inquisição de Lisboa para treinar o modelo de HTR *Portuguese Handwriting 16th-19th c.*[13]. Este apresenta um CER de 5,2%. É de uso público desde setembro de 2023 na plataforma Transkribus[14], e os dados encontram-se em acesso aberto no repositório Zenodo[15].

[10] A taxa de CER é obtida através da comparação feita pela máquina entre o texto de previsão (texto transcrito automaticamente por HTR) e o texto de referência (texto final após correção dos erros no texto de previsão, tendo estatuto de GT). É acompanhada por outra taxa de erro, o *Word Error Rate* (WER), em geral quatro vezes superior ao CER.

[11] URL: https://novaresearch.unl.pt/en/publications/les-archives-inquisitoriales-portugal-sous-htr-le-projet-traprinq/activities/

[12] Carla Vieira, Hervé Baudry (responsável do projeto), Jorge Ferreira Paulo, Leonor Dias Garcia, Maria Olinda Pereira, Margarida Dias da Silva, Mário Soares, Marize Helena de Campos, Natalia Casagrande Salvador, Susana Tavares Pedro (vice-responsável do projeto) e Suzana Severs.

[13] Projeto TraPrInq. Descrição do modelo: "Generic Model created in the framework of the TraPrInq Project (01.2022 to 07.2023) funded by the FCT (Portuguese Agency for Scientific Research), by the members of the team: Carla Vieira, Jorge Ferreira Paulo, Hervé Baudry, Leonor Dias Garcia, Ana Margarida Dias da Silva, Maria Olinda Alves Pereira, Mário Soares Fatela, Marize Helena de Campos, Natalia Casagrande Salvador, Susana Tavares Pedro, Suzana Maria de Sousa Santos Severs. This HTR-model is based on the trial records of the Portuguese Inquisition produced between 1536 (some documents even before) and 1821. It contains careful transcription from 6226 pages (Validation Set: 505 p; Training Set: 5721 p) extracted from 830 processes, mainly by the Lisbon court, with a total of 1268040 words (VS: 107760 words; TS: 1160280). Digitized files can be found on the website of the Portuguese National Archive (Arquivo Nacional da Torre do Tombo). The Model proved its efficacy with hybrid texts (fill-in forms), documents from non-inquisitorial areas. In broad, the transcription reproduces the spelling of words and abbreviations, uses special characters for baseline abbreviation signs and a single COMBINING MACRON for all superscript abbreviation signs, and modernises word separation. The detailed transcription protocol and character list are available at: https://site-2011948.mozfiles.com/files/2011948/Grelha_Criterios.pdf".

[14] URL: https://readcoop.eu/model/portuguese-handwriting-16th-19th-century/

[15] URL: https://zenodo.org/records/13986218/

O modelo foi fruto de um trabalho colaborativo. A produção individual de dados (GT), após revisão, era partilhada em coleções que, ao fim do sexto mês, abriram a via para o primeiro treino do modelo[16]. O HTR, para agilizar os trabalhos, foi possível a partir do terceiro treino do modelo (com mais de 400 000 palavras), os transcritores, produzindo GT, tornaram-se, mais ou menos frequentemente, corretores das transcrições feitas pela máquina. Os trabalhos foram concluídos em julho de 2023, providenciando a última versão do modelo após o nono treino (1 268 040 palavras em 6226 páginas, dados treinados durante sete dias).

Desde setembro de 2023 que o modelo de IA *Portuguese Handwriting 16th-19th c.*, assim como os dados, são acessíveis a qualquer um. O modelo dá resultados com taxas de exatidão variadas, conforme os materiais. Os artigos publicados neste número de *Cultura – Revista de História e Teoria de Ideias* aludem a um ou outro aspeto dos trabalhos e das funcionalidades acima mencionadas, nomeadamente o reconhecimento automático de texto. Mostram que, entre limitações e novos horizontes, o HTR é um dos meios mais potentes e prometedores nas Humanidades Digitais e, de um ponto de vista geral, nos estudos históricos e patrimoniais.

Estudos Inquisitoriais e Humanidades Digitais

Os Estudos Inquisitoriais começaram a dialogar com as Humanidades Digitais no fim dos anos 80 do século XX. Uma nota de rodapé num livro publicado em 1986 alude à elaboração, em curso, de uma base de dados computadorizada (Henningsen 1986, 4-5 n. 7). O ano de referência é 1987. De 17 a 20 de fevereiro, teve lugar um colóquio luso-brasileiro sobre a Inquisição. Três das 106 sessões de trabalho tratavam de informática, apresentando projetos em curso. O investigador responsável, o sociólogo Robert Rowland, concluiu nesses termos:

> A informática pode, hoje, facilitar essa tarefa de reconstituição, permitindo-nos a utilização da documentação inquisitorial como fonte para a história social, mas com uma condição fundamental: a de o computador, e os respectivos programas, serem encarados apenas como ferramentas a utilizar pelos historiadores na prática do seu oficio. (Rowland 1990, 1565)

[16] Sobre os treinos, ver *eInquisição* (URL: https://traprinq.hypotheses.org/).

O projeto apresentado tratava dos processos da Inquisição de Lisboa (aprox. 18 000) (Bethencourt et al. 1990, 1516), cruzando história social, demografia e antropologia. Criou-se uma base de dados através do levantamento manual de processos do século XVI (mais de 4250) e organizada em 27 campos. O objetivo do inventário eletrónico era facilitar as investigações dentro dos processos.

De 7 a 10 de outubro do mesmo ano, foi a vez do congresso "Judicial records and the Computer: inquisitorial, ecclesiastical and secular courts in modern Europe (Bordeaux, 7-10 October 1987)"[17]. O encontro científico, organizado no quadro da European Science Foundation (ESF, Florença), era dedicado ao "diálogo" entre a história e a informática. Encontramos alguns dos participantes do colóquio de fevereiro, como Charles Amiel, Francisco Bethencourt e Robert Rowland, ao lado de Jean-Pierre Dedieu e Jesús Martínez De Bujanda. Todos eles contavam-se entre os mais influentes historiadores das Inquisições. Lê-se poucos anos mais tarde:

> Os planos de Robert Rowland para informatizar os registos da Inquisição portuguesa facilitarão a realização de estudos monográficos, que esperamos venham a lançar uma nova luz sobre a normatividade e a prática da censura. (De Bujanda 1995, 18)[18]

Encerraremos este breve mergulho nos tempos fundadores, citando uma passagem do relatório final do congresso de Bordéus. Redigido por Dedieu, faz-nos tomar, em 2025, vertiginosamente a medida da evolução das coisas:

> Um disco rígido é essencial. Uma disquete de grande capacidade pode ser útil. Verificar se as disquetes são legíveis por máquinas comuns. Uma impressora, mesmo uma básica, é essencial. Um monitor a cores é desnecessário na maioria dos casos. (Dedieu 1987, 10)[19]

Entre os anos 90, a década do "insaciável apetite" e da "espécie de embriaguez na qual tudo parecia possível" (Rowland 1991, 373-374), e 2020, a Revolução Digital atingiu todos os setores, da investigação científica, no seu canto, às sociedades humanas e à vida quotidiana. Assim, foram produzidas e amplificadas bases de dados, como a de Charles Amiel finalizada por Bruno Feitler sobre a Inquisição de Goa, baseada no Reportório de João Delgado Figueira (1623)[20]. Incluindo a Inquisição Portuguesa, aguarda-se a Early Modern

[17] Historical Archives of the EU, ESF 1202. URL: https://archives.eui.eu/en/fonds/475682?item=ESF-1202/
[18] Tradução dos autores.
[19] Tradução dos autores.
[20] Biblioteca Nacional de Portugal, Cod. 203.

Inquisition Database (EMID), base de dados "*trial-centric*" dirigida por Gunnar Knutsen. Outro marco, já um pouco remoto, mas decisivo quanto aos trabalhos apresentados aqui: nos anos 2007-2009, procedeu-se à digitalização do subfundo dos processos da Inquisição de Lisboa[21]. Em princípio, todos, uma vez digitalizados e em boa condição material, se tornaram alvo do HTR. O que parecia impensável até há pouco tempo tornou-se, não só pensável, mas possível.

O primeiro modelo público registado na plataforma Transkribus remonta a março de 2019[22]. Foi o ano em que se treinou o primeiro para o português. A lista dos modelos para esta língua dada mais abaixo[23] (**Fig.2**) é ordenada pela data de disponibilização ao público; são também indicados: o número de palavras treinadas e o CER no Validation Set[24]. Exceto o primeiro modelo, inicialmente treinado no motor HTR+ que deixou de ser utilizado por Transkribus em 2022, todos utilizam o motor PyLaia.

	Nome do modelo	Data de publicação	N.º de palavras	CER
1.	Latin Portuguese Print 17th century[25]	julho 2019 (HTR+)	23 363	1,50%
2.	Transkribus Print M1[26]	fevereiro 2022	5 068 310	2,20%
3.	General Portuguese M1[27]	setembro 2022	64 842	3,80%

[21] URL: https://antt.dglab.gov.pt/exposicoes-virtuais-2/inquisicao-de-lisboa-online/

[22] Mais exatamente, foram dois, para impressos em espanhol. Treinados por Stefano Bazzaco: GothicEsp1500_1 e GOTHIC_BN (Bazzaco 2022).

[23] Língua única nos modelos, exceto os n.º 1 e 5, que são multilíngues (ver a descrição dos modelos).

[24] Para se conhecer a robustez de um modelo, os dados são divididos em dois *sets:* um com um máximo de 10% dos dados *(Validation Set)*, o outro com todo o resto *(Training Set)*. Os dados do primeiro não são treinados, mas servem para a máquina efetuar uma transcrição automática, comparando no fim o texto obtido com o GT.

[25] Hervé Baudry. Descrição do modelo: "This model is based on the Index of censorship printed by Pedro Craesbeeck, a key Lisbon printer of the early seventeenth century. It has been carried out as part of the research project 'The relevance of book expurgation in the procedures of the Portuguese Inquisition (1536-1821): a systematic and individualized approach', realised with the support of CHAM (NOVA FCSH/UAc) through the strategic project sponsored by the National Portuguese Agency for Research (FCT, UIDB/04666/2020)".

[26] Transkribus Community (modelo para 16 línguas). Descrição do modelo: "Extended multi-language Transkribus print model, including antiqua and blackletter prints, typewriter, computer print outs and decorative fonts Includes more languages than print 0.3. The CER in M1 is higher than in 0.3 which is due to a more varied validation set. For languages that were already included in 0.3 the new M1 usually performs equally well or slightly better than 0.3. Curated by the Transkribus team, this model is occasionally updated with community data for continuous improvement".

[27] Lucia Xavier. Descrição do modelo: "First attempt to create a general Portuguese model. The ground truth consists of handwritten and printed documents. Some documents are damaged. This project is a collaboration between two different projects". Ver Lose 2024; Magalhães 2021.

	Nome do modelo	Data de publicação	N.º de palavras	CER
4.	Transkribus portuguese handwriting[28]	outubro 2022	707 803	8%
5.	SPJCL 17C 4.2 Portuguese[29]	junho 2023	64 324	5,60%
6.	Portuguese Handwriting 16th-19th c.	setembro 2023	1 268 040	5,20%
7.	XXth century Typewritten Portuguese[30]	novembro 2023	7468	2,60%
8.	Early Portuguese Printing[31]	outubro 2024	122 754	2,67%

2 Lista dos modelos de IA para a transcrição automática de manuscritos e/ou impressos em português.

[28] Transkribus Community. Descrição do modelo: "General model for Portuguese handwriting. Curated by the Transkribus team, this model is occasionally updated with community data for continuous improvement". Os modelos 2 e 4 foram treinados com dados disponíveis na plataforma.

[29] Descrição do modelo: "Late 17th century Portuguese and Spanish handwriting with some Romanized Hebrew words, based on the Annual Accounts of the Spanish & Portuguese Jews Congregation London 5436-5441 1676-1681".

[30] Natália Salvador. Descrição do modelo: "This model is created from Portuguese typewritten transcriptions from the mid-XXth century, which were based on a confraternity Statute from the XVIIIth century. A great number of documents from Minas Gerais in the XVIIIth century have been transcribed by researchers from the SPHAN (Serviço do Patrimônio Histórico e Artístico Nacional 1936-1970). These transcriptions were then typewritten and are now available digitally in IPHAN´s archives. This model allows a fast and reliable recognition of these types of documents. For this model we followed as much as possible the exact characters and spacing in the original (even when there was lapsus calamus), in order to teach the model to read exactly what is there. The diacritics have been maintained as in the original, except when faded or were too far away from the letter, in those cases we ignored them. Comas, stop points, and others, have remained exactly where they are shown. The markings of a line change, although sometimes appear at the bottom of the word, have been standardized after the last word of the line. Letters that are absent or too faded, have been ignored, leaving a space where they should be".

[31] Saulo Rogério Pacheco Rocha. Descrição do modelo: "This model was trained on a dataset of selected Portuguese grammars and linguistic publications spanning the 16th to the 18th centuries. These documents, along with many others, are publicly accessible through the Portuguese National Digital Library (bndigital.bnportugal.gov.pt). The training set for this version comprises 122,754 words (676 pages) printed in Portuguese since 1536. The dataset reveals texts that include unique letters, diacritics, historical acronyms, typography, and fleurons characteristic of the historical Portuguese writing system adapted to the new press technology, all of which this model has been trained to recognize. Given the linguistic focus, both grammatical and historical, of its training set, this model can also recognize certain Greek letters, Latin text, table patterns and simple initial capitals. However, due to the limited training in these areas, it is not recommended for those uses. This model was developed as part of a master's degree project in the postgraduate linguistics program at the Universidade Federal de Santa Catarina (UFSC). The author was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES)".

Os estudos que formam este número contribuem para as Humanidades Digitais e os estudos inquisitoriais, alguns deles contemplando frontalmente o uso das tecnologias do reconhecimento de texto. O artigo de Luís Henrique Menezes Fernandes debruça-se sobre a revitalização digital da Bíblia Almeida, que foi a primeira tradução do Novo Testamento em português (século XVII) por João Ferreira Almeida, convertido ao protestantismo. Este estudo que, além do interesse filológico e histórico, abre os horizontes da cultura portuguesa no Oriente na sua vertente heterodoxa. Em "Limites e possibilidades do modelo de HTR *Portuguese Handwriting 16th-19th c.*", Susana Tavares Pedro abre caminho à reflexão sobre o modelo de Inteligência Artificial agora disponível para todos na plataforma de paleografia digital Transkribus. Com uma saudável intenção propedêutica, a autora previne os interessados contra as falsas esperanças, provando a contrario a eficácia desta ferramenta. O artigo de Ana Margarida Dias da Silva, "Transcrições automáticas nos arquivos distritais portugueses: acelerar o acesso à informação", prova a importância e o potencial do recurso aos meios da transcrição automática para os arquivos. Os trabalhos, que foram conduzidos em fundos não-inquisitoriais, vêm confirmar a eficácia do modelo treinado pela equipa do projeto TraPrInq. A exploração de milhares de páginas de processos da Inquisição de Lisboa tornou visível um fenómeno até agora pouco assinalado, a presença de documentos impressos. Quantitativamente, ainda se trata de uma minoria, mas apresentam aspetos múltiplos de grande interesse como o mostram os autores de "Documentos híbridos e HTR: os impressos nos processos da Inquisição portuguesa", Hervé Baudry e Natália Salvador.

Bibliografia

ALKENDI, Wissam, Franck Gechter, Laurent Heyberger, e Christophe Guyeux. 2024. "Advancements and Challenges in Handwritten Text Recognition: A Comprehensive Survey". *Journal of Imaging* 10 (1): 18. https://doi.org/10.3390/jimaging10010018.

ALVES, Daniel. 2016. "As Humanidades Digitais como uma comunidade de práticas dentro do formalismo académico: dos exemplos internacionais ao caso português". *Ler História* 69: 91-103. https://doi.org/10.4000/lerhistoria.2496.

BAZZACO, Stefano, Ana Milagros Jiménez Ruiz, Ángela Torralba Ruberte, e Mónica Martín Molares. 2022. "Sistemas de reconocimiento de textos e impresos hispánicos de la Edad Moderna. La creación de unos modelos de HTR para la transcripción automatizada de documentos en gótica y redonda (s. XV-XVII)". *Historias Fingidas* 1 (número especial: *Humanidades Digitales y estudios literarios hispánicos*): 67-125. https://doi.org/10.13136/2284-2667/1190.

BETHENCOURT, Francisco, Piedade Braga Santos, Robert Rowland, e Teresa Rodrigues. 1990. "Informática e Inquisição: Reflexões em Torno de um Projecto". In *Inquisição: comunicações apresentadas ao 1º Congresso Luso-Brasileiro sobre Inquisição*, org. Maria Helena Carvalho dos Santos, vol. 3, 1513-1523. Lisboa: Universitária Editora-S.P.E. Séc. XVIII.

CIULA, Arianna. 2005. "Digital palaeography: using the digital representation of medieval script to support palaeographic analysis". *Digital Medievalist* 1. http://doi.org/10.16995/dm.4.

COUTURE, Beatrice, Farah Verret, Maxime Gohier, e Dominique Deslandres. 2022. 2022. "The challenges of HTR model training: Feedbacks from the Project Donner le gout de l'archive a l'ere numerique". *Journal of Data Mining & Digital Humanities*. http://doi.org/10.48550/arXiv.2212.11146.

DE BUJANDA, Jesús Martínez, dir. 1995. *Index de l'Inquisition portugaise, 1547, 1551, 1561, 1564, 1581*. Sherbrooke, Québec: Centre d'études de la Renaissance, Université de Sherbrooke.

DEDIEU, Jean Pierre. 1987. "Rapport introductif par Jean Pierre Dedieu: Les archives judiciaires et l'ordinateur". Conference "Judicial records and the Computer: inquisitorial, ecclesiastical and secular courts in modern Europe" (Bordeaux, 07-10/10/1987), European Science Foundation: ESF-1202, 7-13. https://archives.eui.eu/en/fonds/475682.

EHRMANN, Maud, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, e Antoine Doucet. 2021. "Named Entity Recognition and Classification on Historical Documents: A Survey". *ACM Computing Surveys* 56 (2): 1-47. https://doi.org/10.48550/arXiv.2109.11406.

HENNINGSEN, Gustav, e John Tedeschi, eds. 1986. *The Inquisition in Early Modern Times Europe. Studies on Sources and Methods*. Dekalb, IL: Northern Illinois University Press.

HUMPHRIES, Mark, Lianne C. Leddy, Quinn Downton, Meredith Legace, John McConnell, Isabella Murray, e Elizabeth Spence. 2024. "Unlocking the Archives: Using Large Language Models to Transcribe Handwritten Historical Documents". *arXiv preprint arXiv:2411.03340*. https://doi.org/10.48550/arXiv.2411.03340.

KIESSLING, Benjamin, Robin Tissot, Peter A. Stokes, e Stökl Daniel Ben Ezra. 2019. "eScriptorium: An Open Source Platform for Historical Document Analysis". In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. Sydney: IEEE. https://doi.org/10.1109/ICDARW.2019.10032.

LANG, Sabine. 2025. "Machine Learning Meets Provenance Research: Recognising and Transcribing Handwritten Annotations in Auction Catalogues". *International Journal of Humanities and Arts Computing* 19 (1): 17-32.

LOSE, Alícia Duha, João Guilherme Andrade dos Santos, Leonardo Coelho Marques de Jesus, Lívia Borges Souza Magalhães, e Lucia Werneck Xavier. 2024. "Transkribus: uma ferramenta de paleografia digital mediando pesquisas em fontes inquisitoriais". *Revista LaborHistórico* 10 (1). https://doi.org/10.24206/ lh.v10i1.63285.

MAGALHÃES, Lívia, e Lucia Werneck Xavier. 2021. "'Can machines think?': Por uma paleografia digital para textos em língua portuguesa". In *Paleografia e suas interfaces*, organizado por Alícia Duah Lose e Arivaldo Sacramento de Souza, 259-269. Salvador: Memória & Arte.

MÜHLBERGER, Günter, Louise Seaward, Melissa Terras, Sofia Ares Oliveira, Vicente Bosch, et al. 2019. "Transforming scholarship in the archives through Handwritten Text Recognition. Transkribus as a case study". *Journal of Documentation - Emerald Publishing* 75 (5): 954-976.

NOCKELS, Joe, Paul Gooding, Sarah Ames, e Melissa Terras. 2022. "Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research". *Archival Science* 22: 367-392. https://doi.org/10.1007/s10502-022-09397-0.

RABUS, Achim. 2019. "Training Generic Models for Handwritten Text Recognition using Transkribus: Opportunities and Pitfalls". In *Proceedings of the Dark Archives Conference*. Acedido a 7/5/2023. https://www.academia.edu/49356690/Training_generic_models_for_Handwritten_Text_Recognition_using_Transkribus_Opportunities_and_pitfalls.

ROLING, Marco. 2020. "Does Handwriting Text Recognition work for damaged archives?". ResearchGate.net. Acedido a 10/5/2023. https://www.researchgate.net/publication/340117708_Does_Handwriting_Text_Recognition_Work_for_Damaged_Archives.

ROMEIN, Annemieke. 2020. "Entangled Histories: OCR + HTR = ATR: Automatic Text Recognition". *KB LAB*. Acedido a 25/5/2023. https://lab.kb.nl/about-us/blog/entangled-histories-ocr-htr-atr-automatic-text-recognition.

ROWLAND, Robert. 1990. "Informática e Inquisição". In *Inquisição: comunicações apresentadas ao 1º Congresso Luso-Brasileiro sobre Inquisição*, org. Maria Helena Carvalho dos Santos, vol. 3, 1561-1565. Lisboa: Universitária Editora-S.P.E. Séc. XVIII.

ROWLAND, Robert. 1991. "Un'esperienza di informatizzazzione dei registri dell'Inquisizione portoghese". In *L'Inquisizione romana in Italia nell'età moderna. Archivi, problemi di metodo e nuove ricerche. Atti del seminario internazionale, Trieste, 18-20 maggio 1988*, 369-390. Roma: Ministerio per i beni culturali e ambientali.

SANTOS RUIZ, Víctor de. 2017. "Paleografía digital: reto y necesidad de los profesionales de archivo". Tese de Mestrado, Universidad Carlos III de Madrid. https://www.lhistoire.fr/irht-dans-le-secret-des-manuscrits/pal%C3%A9ographie-la-r%C3%A9volution-num%C3%A9rique.

SEFIL, Kutay. 2024. *L'implémentation de l'OCR dans une bibliothèque patrimoniale. L'exemple de la Bibliothèque interuniversitaire de la Sorbonne*. Mémoire de Master. Paris: École nationale des chartes.

STUTZMANN, Dominique. 2017. "Paléographie: la révolution numérique". *L'Histoire* 439: 30-31.

TARTE Segolene, Tal Hassner, Robert Sablatnig, e Dominique Stutzmann. 2014. "Digital Palaeography: New Machines and Old Texts (Dagstuhl Seminar 14302)". *Dagstuhl Reports* 4 (7): 112-135. https://doi.org/10.4230/DagRep.4.7.1127.

WANG, Haifeng, Changzai Pan, Xiao Guo, Chunlin Ji, e Ke Deng. 2021. "From object detection to text detection and recognition: A brief evolution history of optical character recognition". *Wiley Interdisciplinary Reviews: Computational Statistics* 13 (5): e1547. https://doi.org/10.1002/wics.1547/.

WANG, Junmiao. 2023. "A Study of The OCR Development History and Directions of Development". *Highlights in Science, Engineering and Technology* 72: 409-415.