CENTRO UNIVERSITÁRIO DE BRASÍLIA
FACULDADE DE TECNOLOGIA E CIÊNCIAS SOCIAIS APLICADAS

RESUMO EXECUTIVO

HistOCR

Membros do Projeto

22151763 | Anna Rafaella Frazão de Oliveira

22151867 | Lucca Resende da Costa Paiva

22154907 | Wendy Nicole Passos da Silva

Orientador

Prof. MSc. Fabiano Mariath D'Oliveira

BRASÍLIA, junho de 2025

AGRADECIMENTOS

Ao professor Fabiano Mariath D'Oliveira, pelo apoio e orientação ao longo do projeto, desde de sua concepção como trabalho de disciplina em Sistemas Distribuídos, e ao Arquivo Público do Distrito Federal, pelo suporte e confiança em compartilhar conosco suas necessidades.

RESUMO

O projeto HistOCR consiste no planejamento e no desenvolvimento de uma solução digital voltada à transcrição automatizada de documentos históricos por meio de tecnologias de reconhecimento óptico de caracteres (OCR), com o objetivo de auxiliar bibliotecários, arquivistas, pesquisadores e instituições culturais na preservação, digitalização e acesso a acervos históricos.

Palavras-chave: OCR, digitalização de documentos históricos, preservação digital

SUMÁRIO

1. PROBLEMA/OPORTUNIDADE **3**
2. BENEFÍCIOS DA SOLUÇÃO **4**
3. PÚBLICO-ALVO **4**
4. PROTÓTIPO VISUAL **4**
5. CONSIDERAÇÕES FINAIS **4**
REFERÊNCIAS **4**

1. PROBLEMA/OPORTUNIDADE

Em 2023, o Cetic.br estimou que o Brasil possui cerca de 14.919 instituições culturais, como museus, bibliotecas e teatros. Um dos principais desafios enfrentados por esses locais é a conservação de artefatos antigos, sujeitos à degradação natural e a desastres, como o incêndio do Museu Nacional. Cerca de 46% das coleções afetadas nesse evento foram perdidas. Nesse contexto, a digitalização surge como alternativa segura, garantindo preservação e maior visibilidade dos acervos culturais em um mundo cada vez mais digital.

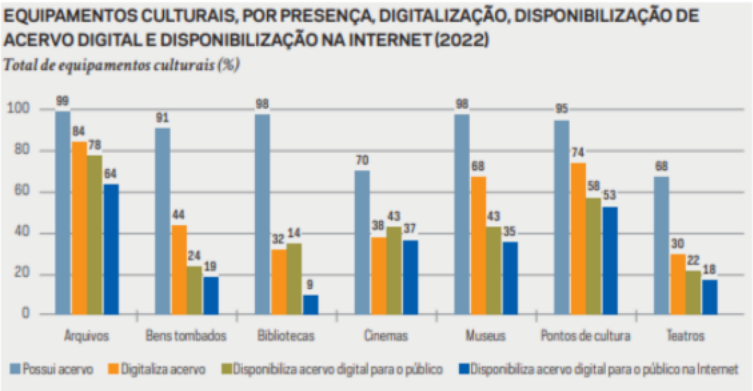

Figura 1 - Gráfico da Digitalização de Equipamentos Culturais no Brasil

A pesquisa indica que a digitalização ainda é minoritária nas instituições culturais, principalmente pela falta de financiamento. Há também uma lacuna tecnológica: OCR tradicional falha em documentos históricos devido a baixa qualidade de imagens e degradação dos materiais.

Fatores linguísticos, como variação e evolução do idioma, exigem modelos específicos que a maioria dos sistemas não possui. Esses desafios abrem uma oportunidade para um aplicativo de transcrição automatizada voltado a acervos históricos. A ferramenta reduziria custos e tempo, beneficiando bibliotecários, arquivistas, pesquisadores e entusiastas. Com alta acurácia e interface intuitiva, o app pode revolucionar a preservação, catalogação e acesso a documentos históricos.

2. BENEFÍCIOS DA SOLUÇÃO

A digitalização de documentos históricos contribui para a preservação e segurança do patrimônio documental, atendendo diretrizes governamentais e promovendo inclusão cultural. Ao automatizar a transcrição, reduz a necessidade de mão de obra, diminui custos operacionais e permite a ampliação de acervos digitais com menor investimento por parte de instituições culturais e acadêmicas.

Além disso, facilita o acesso remoto e pesquisável a documentos históricos, democratizando a informação para pesquisadores, estudantes e o público em geral. A adoção de tecnologias modernas, como o aprendizado de máquina, ainda abre espaço para melhorias contínuas na acurácia e aplicabilidade em diferentes contextos históricos e linguísticos.

3. PÚBLICO-ALVO

Bibliotecários, arquivistas, pesquisadores e instituições como museus e arquivos enfrentam desafios significativos na preservação, digitalização e acesso a documentos históricos. A morosidade da digitalização manual, a deterioração dos acervos e a ausência de sistemas eficientes para busca e indexação comprometem a integridade e a disponibilidade das informações. Além disso, pesquisadores lidam com acesso limitado, dificuldade na leitura de manuscritos antigos e tempo elevado gasto com transcrições manuais.

A solução proposta no projeto contribui para superar essas barreiras por meio de funcionalidades como digitalização e transcrição automatizadas, indexação inteligente, acesso remoto e preservação digital. Essas melhorias reduzem custos operacionais, otimizam o tempo de trabalho e ampliam o alcance dos acervos históricos, promovendo maior visibilidade e democratização do conhecimento.

4. PROTÓTIPO VISUAL

O protótipo visual foi construído utilizando a ferramenta Figma. Abaixo estão listados alguns conteúdos e seus respectivos fluxos em imagens, que seguem a sequência da esquerda para a direita. Para visualizar mais, acesse o link: Figma/HistOCR.

Fluxo de Login de Usuário:

Ao entrar no aplicativo, o usuário consegue realizar seu login utilizando uma conta Google.

Em seguida, para completar seu cadastro, é solicitado a profissão do usuário, que pode ser Arquivista, Bibliotecário(a), Historiador(a), Museólogo(a) ou Outro. Caso o usuário seja um Estudante, por exemplo, ele deverá selecionar a opção Outro e em seguida preencher escrevendo "Estudante" no campo adicional que irá aparecer. No caso do fragmento do protótipo acima, o usuário é um Arquivista e, portanto, não aparecerá o campo adicional para ser preenchido.

Fluxo de Transcrição de Documento:

Após o login, o usuário acessa a tela inicial com as últimas transcrições, opções de chat e configurações. Ao iniciar o chat, pode enviar uma imagem ou tirar foto para transcrição, mediante permissão. O texto gerado pode ser avaliado e corrigido manualmente, ajudando o modelo a melhorar seu desempenho.

Fluxo de Transcrição de Documento:

Ao escolher corrigir o texto manualmente, o usuário visualiza um modal com a imagem e um campo editável. Essa etapa é opcional; se recusada, o app prossegue para a finalização da transcrição. O modelo sugere um título, mas o usuário pode editá-lo clicando no ícone de lápis.

Fluxo de Correção Manual do texto e Modificação do Nome do arquivo :

Ao retornar à tela inicial, o usuário pode acessar todas as transcrições salvas ou buscar por nome. O histórico exibe repositórios individuais (privados) e de organização (compartilhados). Na visualização do documento, é possível ver a imagem original, o texto transcrito, editar o título e copiar o conteúdo.

5. CONSIDERAÇÕES FINAIS

O desenvolvimento do HistOCR permitiu alcançar com êxito os objetivos propostos, cumprindo com seus requisitos e entregando uma solução funcional, acessível e inovadora para a transcrição automatizada de documentos históricos. Além disso,o reconhecimento pelo Arquivo Público do Distrito Federal proporcionou uma parceria fundamental para validar a aplicabilidade da solução em um contexto real, ampliando seu impacto social e institucional.

REFERÊNCIAS

AFFINDA. **From OCR to AI: The Evolution of OCR Technology**. 14 mar. 2024. Disponível em: https://www.affinda.com/blog/from-ocr-to-ai-the-evolution-of-ocr-technology. Acesso em: 6 maio 2025.

AGÊNCIA BRASIL. **Preservação e digitalização de acervos é desafio, dizem especialistas. ISTOÉ Dinheiro, 13 nov. 2018**. Disponível em: https://istoedinheiro.com.br/preservacao-e-digitalizacao-de-acervos-e-desafio-dizem-especialistas. Acesso em: 6 maio 2025.

BRASIL**. Arquivo Nacional. Digitalização de acervo no Arquivo Nacional. Rio de Janeiro**: Governo Federal, 2024. Disponível em: https://www.gov.br/arquivonacional/pt-br/canais_atendimento/imprensa/copy_of_noticias/digitalizacao-de-acervo-no-arquivo-nacional. Acesso em: 10 abr. 2025.

BRASIL. **Ministério da Ciência, Tecnologia e Inovação. Governo Federal assina decreto de incentivo a projetos tecnológicos de alto impacto. Gov.br**, Brasília, 10 jun. 2024. Disponível em: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/noticias/2024/06/governo-federal-assina-decreto-de-incentivo-a-projetos-tecnologicos-de-alto-impacto. Acesso em: 6 maio 2025.

BRASIL**. Ministério da Ciência, Tecnologia e Inovação. Mais Inovação vai investir R$ 66 bilhões em projetos até 2026. Brasília: MCTI,** 16 jan. 2024. Disponível em: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/noticias/2024/01/mais-inovacao-vai-investir-r-66-bilhoes-em-projetos-ate-2026. Acesso em: 6 maio 2025.

BRASIL. **Ministério da Cultura. Sistema Nacional de Bibliotecas Públicas (SNBP) – Informações das bibliotecas públicas. Gov.br**, [s.l.], 25 fev. 2022 (atualizado em 24 mar. 2022). Disponível em: https://www.gov.br/cultura/pt-br/assuntos/sistema-nacional-de-bibliotecas-publicas-snbp/teste01/informacoes-das-bibliotecas-publicas-1. Acesso em: 07 jun. 2025.

BRASIL. **Portal da Transparência. Função 13 – Cultura. Brasília**, [s.d.]. Disponível em: https://portaldatransparencia.gov.br/funcoes/13-cultura. Acesso em: 07 jun. 2025.

CETIC.BR – CENTRO REGIONAL DE ESTUDOS PARA O DESENVOLVIMENTO DA SOCIEDADE DA INFORMAÇÃO**. Pesquisa sobre o uso das tecnologias de informação e comunicação nos equipamentos culturais brasileiros: TIC Cultura** *2022.* 1. ed. São Paulo: Comitê Gestor da Internet no Brasil, 2023. ISBN: 978-65-86949-98-8. Disponível em: https://cetic.br/media/docs/publicacoes/2/20230621154638/tic_cultura_2022_livro_eletronico.pdf. Acesso em: 17 mar. 2025.

CGI.br/NIC.br. **TIC Cultura 2022: Equipamentos culturais, por dificuldades de digitalização do acervo**. São Paulo: Cetic.br, 2022. Disponível em: https://cetic.br/pt/tics/cultura/2022/geral/D6/. Acesso em: 6 maio 2025.

CORDELL, Ryan Charles. **Machine Learning + Libraries: A Report on the State of the Field. 2020. LC Labs, Library of Congress.** Disponível em: https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf?loclr=blogsig. Acesso em: 8 mar. 2025.

FERREIRA, Rodrigo Santos; NASCIMENTO, Carolina Rocha. **Uso de indicadores da qualidade na avaliação do desempenho de hospitais públicos. Cadernos de Saúde Pública**, Rio de Janeiro, v. 37, n. 4, e00151420, 2021. Disponível em: https://doi.org/10.1590/0102-311X00151420. Acesso em: 10 abr. 2025.

GULLINO, Daniel; ÉBOLI, Evandro. **Iphan tem em 2021 o menor orçamento dos últimos 10 anos. O Globo**, 15 ago. 2021. Disponível em: https://oglobo.globo.com/brasil/iphan-tem-em-2021-menor-orcamento-dos-ultimos-10-anos-25156053. Acesso em: 6 maio 2025.

ISO/IEC**. ISO/IEC 25010:2011 – Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models. International Organization for Standardization**, 2011.

METELING, Michael. **Optical Character Recognition (OCR): What Is OCR, Its Benefits, Limitations, and AI. Interscan LLC, 31 out. 2023**. Disponível em: https://www.interscanllc.com/blog/optical-character-recognition-ocr. Acesso em: 6 maio 2025.

ODA, Rafael; BAHIA, Eliana Maria dos Santos. **A evolução dos arquivos públicos no cenário brasileiro (2008–2022). TransInformação, Campinas**, v. 36, e2410349, 2024. Disponível em: https://doi.org/10.1590/2318-0889202436e2410349. Acesso em: 10 abr. 2025.

PONCELAS, Alberto et al. ***A Tool for Facilitating OCR Postediting in Historical Documents.*** **ADAPT Centre, Dublin City University; Trinity College Dublin.** Disponível em: https://aclanthology.org/2020.lt4hala-1.7.pdf. Acesso em: 8 mar. 2025.

PRESSMAN, Roger S.; MAXIM, Bruce R. **Engenharia de Software: Uma Abordagem Profissional. 8. ed.** Porto Alegre: McGraw Hill, 2016.

SANTOS, Gildenir Carolino; GRÁCIO, José Carlo Abbud. **Políticas de preservação digital no Brasil: um panorama do estado da arte. Revista Brasileira de Preservação Digital, Campinas**, v. 5, 2024. Disponível em: https://econtents.bc.unicamp.br/inpec/index.php/rebpred/article/view/20185. Acesso em: 6 maio 2025.

SANTOS, Maria Clara. **A preservação dos documentos históricos em ambientes digitais. Revista Brasileira de Preservação Digital, Campinas**, v. 5, n. 2, p. 45–60, 2021. Disponível em: https://econtents.bc.unicamp.br/inpec/index.php/rebpred/article/view/13858. Acesso em: 6 maio 2025.

SILVA, Carlos Alberto Ávila; MOURA, Maria das Graças Targino. ***Desafios e avanços na recuperação automática da informação.*** **Ciência da Informação, Brasília**, v. 49, n. 2, p. 1–12, maio/ago. 2020. Disponível em: https://www.scielo.br/j/ci/a/3RKGt6c6RY4pCFYWcypKh5B/. Acesso em: 6 maio 2025.

SILVA, José Carlos. **O papel da sociedade na preservação de bens históricos e culturais. Consultor Jurídico, São Paulo,** 9 fev. 2025. Disponível em: https://www.conjur.com.br/2025-fev-09/o-papel-da-sociedade-na-preservacao-de-bens-historicos-e-culturais/. Acesso em: 6 maio 2025.

SOMMERVILLE, Ian**. Engenharia de Software. 10**. ed. São Paulo: Pearson, 2019.

TREVO SOLUÇÕES EM COMUNICAÇÃO – ASSESSORIA DE COMUNICAÇÃO DO MUSEU NACIONAL. **Museu Nacional apresenta balanços após um ano do incêndio. Rio de Janeiro: Museu Nacional,** 2019. Disponível em: https://www.museunacional.ufrj.br/destaques/balan%C3%A7o_resgatehtml.html. Acesso em: 10 mar. 2025.