

Memorial Digital Coletivo Automatizado: um fluxo de trabalho com IA para curadoria de dados

Luís Flávio Ferreira Monteiro, Cristiano Maciel

Instituto de Computação – Universidade Federal de Mato Grosso (UFMT)
Cuiabá – MT – Brazil

luisflaviomont@gmail.com, cristiano.maciel@ufmt.br

Abstract. Introduction: Generating collective digital memorials from public web data requires workflows that balance speed, privacy, and cultural sensitivity. **Objective:** To propose a four-step workflow for creating collective digital memorials using publicly available web data. **Steps:** Scrape sources, clean extracted content, classify items with the local Gemma 3-1B model, and produce automatic summaries. **Results:** In tests with two real profiles, assembly time fell from about 25 minutes to 1.6 minutes (93% reduction), and section-based organization reached around 85% accuracy. The system runs locally, preserves privacy, and requires brief human review to ensure cultural sensitivity. Limitations include a restricted test set and dependence on a compact model. Future work will explore larger models, multimodal support, and evaluations with grieving families.

Keywords Digital Legacy, Death, Artificial Intelligence, Digital Memorial, Collective, Data Curation

Resumo. Introdução: A montagem de memoriais digitais coletivos a partir de dados públicos exige fluxos de trabalho que equilibrem velocidade, privacidade e sensibilidade cultural. **Objetivo:** Propor um fluxo de trabalho de quatro etapas para gerar memoriais digitais coletivos a partir de dados disponíveis na web. **Etapas:** raspagem das fontes, limpeza do conteúdo, classificação dos itens com o modelo Gemma 3-1B executado localmente e produção de resumos automáticos. **Resultados:** em testes com dois perfis reais, o tempo de montagem caiu de cerca de 25 min para 1,6 min (redução de 93 %) e a organização por seções alcançou cerca de 85 % de acerto. O sistema opera localmente, preserva a privacidade e exige revisão humana para garantir sensibilidade cultural. As limitações incluem o conjunto de teste restrito e a dependência de um modelo compacto. Trabalhos futuros envolvem o uso de modelos maiores, suporte multimodal e avaliações com familiares enlutados.

Palavras-Chave Legado Digital, Morte, Inteligência Artificial, Memorial Digital, Coletivo, Curadoria de Dados

1. Introdução

Com o aumento do volume de dados pessoais disponíveis online, surge o desafio de preservar essas informações para a posteridade, especialmente após o falecimento de indivíduos [Trevisan et al. 2021]. Estima-se, por exemplo, que o Facebook poderá alcançar cerca de 4,9 bilhões de perfis de usuários falecidos até o ano 2100 [Öhman & Watson 2019]. Esse acervo massivo de dados pessoais levanta questões sobre quem terá

acesso a essas memórias e como elas poderão ser utilizadas de forma respeitosa e historicamente relevante.

A importância de preservar memórias digitais enfrenta diversos desafios atuais. Conteúdos online são voláteis, contas podem ser desativadas ou excluídas, links tornam-se inválidos e políticas de privacidade restringem acesso a informações valiosas. Além disso, a curadoria manual de um memorial digital, como a coleta de posts, fotos e depoimentos, é uma tarefa que requer esforços significativos, especialmente para familiares ou enlutados, devido à complexidade de organizar conteúdos dispersos e à possível escassez de informações relevantes [Trevisan, Maciel & Bim, 2021].

Neste trabalho, propomos um sistema automatizado, baseado em Inteligência Artificial (IA), que gera memoriais digitais coletivos a partir de dados públicos disponíveis na web. Entende-se por memorial digital coletivo um repositório online que agrega e organiza, de forma centralizada, conteúdos representativos das trajetórias ou legados de múltiplas pessoas ou eventos relacionados, promovendo a construção de uma memória compartilhada [Ueda et al. 2022]. Diferentemente de simples perfis memorializados em redes sociais, nosso enfoque está em agregar múltiplas fontes de dados e aplicar Inteligência Artificial (IA) para ajudar na curadoria do conteúdo.

O memorial digital coletivo já existente é uma aplicação web (.NET + HTML/CSS/JavaScript) que oferece cadastro protegido por login, formulário passo-a-passo para preencher cinco seções (História de Vida, Carreira, Religião, Hobbies e Galeria), página de detalhe com biografia, fotos e mural de homenagens. O novo fluxo de trabalho da IA conecta-se a essa plataforma via API REST: em vez de o usuário digitar tudo manualmente, ele só informa o nome do homenageado e links públicos; o sistema gera automaticamente os rascunhos de cada seção e devolve-os pré-preenchidos na interface, prontos para a revisão humana e edição final.

Vale destacar que a proposta dialoga diretamente com os desafios “Implicações da Inteligência Artificial em IHC”, “Ética em IHC” e “Aspectos culturais” do GranDIHC-BR 2025-2035 [Pereira et al. 2024], unindo automação com revisão humana para curar memórias digitais de forma transparente e respeitosa.

2. Metodologia

Este estudo, classificado como pesquisa aplicada e exploratório-descritiva [Gil 2019] e, adota um método misto quantitativo e qualitativo [Creswell & Creswell 2018], combinando dados numéricos e análises qualitativas. O sistema automatizado para memoriais digitais coletivos processa dados públicos em um fluxo contínuo.

Inicialmente, o fluxo de trabalho extrai informações de fontes públicas, como perfis e blogs, respeitando regras de acesso (*robots.txt* e termos de uso) [Brown et al. 2024]. Utilizamos *requests* e *BeautifulSoup* para páginas estáticas [Reitz 2025; Richardson 2025], *Selenium WebDriver* para conteúdo dinâmico [Selenium Developers 2025]. Textos e imagens são salvos em formato estruturado.

Em seguida, os dados são limpos com *pandas* e expressões regulares, eliminando duplicatas e padronizando formatos [McKinney 2010]. O modelo *Sentence-T5-base* seleciona trechos relevantes, transformando o conteúdo em vetores [Ni et al. 2021].

Posteriormente, o modelo *Gemma 3-1B*, via *HuggingFace Transformers* [Wolf et al. 2019], organiza o conteúdo nas seções do sistema de memorial digital concebido pelo grupo [DAVI 2025]: História de Vida, Carreira, Religião, Hobbies e Galeria.

Resumos são gerados, verificados contra fontes originais e, se necessário, marcados como “Conteúdo indisponível”.

Por fim, o memorial é enviado à aplicação web, onde o usuário-curador revisa o conteúdo, assegurando precisão factual e ética. Para avaliar a automação, um quasi-experimento mediu o tempo de processamento de perfis de teste (De um pesquisador e um professor) contra o tempo manual de 25 minutos, que foi obtido por meio de duas fontes, de [Lopes et al. 2014] e do grupo [DAVI 2025], calculando redução percentual e métricas de qualidade (precisão e *F1-score*).

2.1 Considerações éticas em pesquisa

Este estudo utilizou exclusivamente dados públicos disponíveis na web, observando estritamente as regras de acesso (`robots.txt` e termos de uso) e sem coletar informações privadas ou sensíveis. Os dados utilizados foram previamente autorizados pelas pessoas homenageadas para uso neste estudo. Todo conteúdo gerado automaticamente pelo modelo Gemma 3-1B foi submetido à revisão humana pelo curador, de modo a evitar vieses e assegurar conformidade com a privacidade dos indivíduos. Por se tratar de dados públicos autorizados e pelo procedimento de revisão, não foi necessário protocolo de aprovação em Comitê de Ética em Pesquisa, mas o processo respeita os princípios da Lei Geral de Proteção de Dados (LGPD) no que tange à privacidade de usuários vivos.

3. Fluxo do Sistema

A arquitetura do sistema segue um estilo em camadas recomendado por [Sommerville 2015], favorecendo o desacoplamento, reuso e manutenção evolutiva. No centro está um serviço Python que coordena todo o fluxo; Python foi escolhido pelo ecossistema maduro de bibliotecas para *scraping*, NLP (Processamento de Linguagem Natural) e IA. Dentro desse serviço há duas subcamadas principais. A de coleta executa raspagem assíncrona, permitindo varrer dezenas de URLs em paralelo enquanto isola falhas de rede. A de processamento faz tokenização e reconhecimento de entidades, utiliza regex e NLTK para limpeza e armazena tudo em *pandas DataFrames* [Bird et al. 2009], que são tabelas, possibilitando filtragens vetorizadas antes de enviar os dados ao modelo Gemma 3-1B.

O coração do fluxo é o Gemma 3-1B [Gemma Team 2025], carregado localmente via *HuggingFace Transformers*. Esse modelo de 1 bilhão de parâmetro executa em CPU e, no sistema proposto, lida tanto com a classificação temática quanto com a redação de resumos [Gemma Team 2025]. Rodar o modelo localmente preserva privacidade, pois nenhum dado sai do servidor, e permite uso mesmo offline, um ponto-chave para famílias que queiram executar a ferramenta em computadores pessoais.

A curadoria automatizada é guiada por prompts pré-definidos por categoria; incluímos exemplos (*few-shot*) para estabilizar estilo e evitar alucinações. Quando o texto gerado não corresponde às fontes, o sistema registra a falha e pode reexecutar com variantes de prompt ou, se disponível, um modelo alternativo, mantendo robustez sem intervenção manual frequente. Na camada de apresentação, geramos uma página (HTML/CSS/JavaScript) que organiza o memorial em seções. Na Figura 1, apresentamos o fluxo completo de geração do memorial digital coletivo, para ilustrar todo o processo que ocorre desde o *input*, até a geração do memorial.

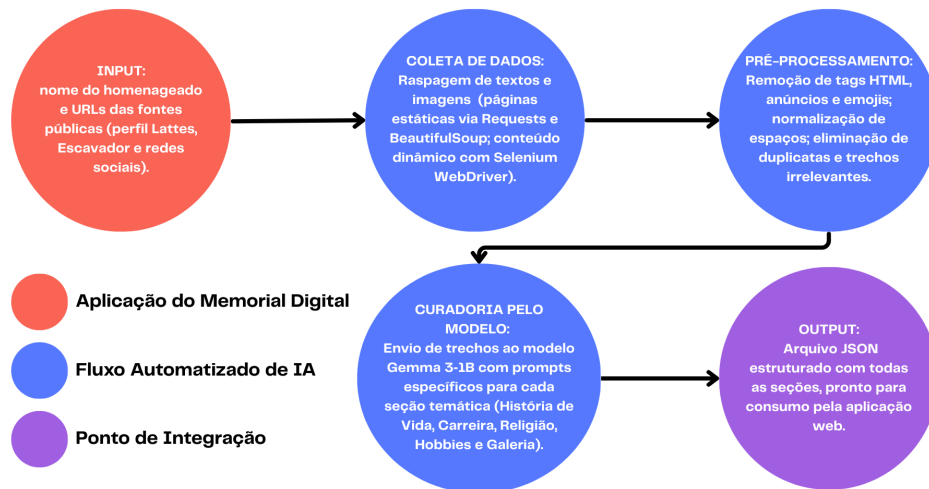


Figura 1: Fluxo completo de geração do memorial digital coletivo

4. Avaliação

Para medir o impacto do fluxo de trabalho, comparamos diretamente dois cenários: criação manual de um memorial (baseline de ~25 min, valor derivado de [Lopes et al. 2014] e testes internos do [DAVI 2025]) e geração automatizada (raspagem + *Gemma 3-1B*). Usamos apenas os perfis público-autorizados do autor e de um professor de uma universidade federal, evitando problemas éticos com o consentimento para uso de dados e controlando variáveis de heterogeneidade.

Nos testes, o tempo médio para criar um memorial manualmente foi em torno de 25 min por memorial. Já o fluxo de trabalho consumiu em média 1,6 minutos (máx. ≈ 2 min para o perfil com mais informações), alcançando uma redução absoluta de 23,4 min e economia de ≈ 93 %. Dessa forma, dez memoriais que antes exigiriam ~4 h de esforço podem ser produzidos em < 30 min de processamento paralelo, um ganho de escalabilidade.

Quanto à qualidade estrutural, avaliamos se cada trecho foi alocado na seção correta tomando a versão manual como referência. O acerto global foi ≈ 85%, com variações por categoria: Carreira ~90%, História de Vida ~85%, Hobbies ~80%. Os erros restantes surgiram em casos limítrofes (ex.: postagem pessoal com conteúdo profissional) e podem ser corrigidos pelo usuário-curador, que interage com a aplicação web para revisar e ajustar o conteúdo, um processo que pode demandar poucos minutos por memorial, mas é essencial para garantir precisão factual e conformidade ética, alinhando-se aos princípios de Interação Humano-Computador.

Esses resultados indicam que o fluxo de trabalho atinge simultaneamente ganho de produtividade e robustez de curadoria. A IA transfere o papel do usuário de “redator” para supervisor, exigindo apenas ajustes pontuais sem comprometer a fidelidade narrativa, premissa fundamental para memoriais digitais reais e escaláveis.

5. Considerações Finais

A preservação de memórias digitais envolve desafios técnicos, sociais e éticos. Demonstramos um fluxo de trabalho com IA que automatiza coleta, categorização e

curadoria, reduzindo mais de 90% do esforço manual sem sacrificar qualidade narrativa. A solução atua como ferramenta de apoio: a IA monta o rascunho; o usuário-curador analisa, ajusta e valida antes da publicação.

O sistema incorpora salvaguardas éticas, processando apenas dados públicos e executando localmente para proteger a privacidade. A revisão humana pelo usuário-curador, realizada em uma interface web, ajusta o conteúdo gerado pelo modelo *Gemma 3-1B*, via *HuggingFace Transformers*, para respeitar sensibilidades culturais, seguindo diretrizes éticas que consideram o contexto dos memoriais. Neste ponto, cabe salientar que a Lei Geral de Proteção de Dados no Brasil, protege dados do usuário em vida, mas não os considera após a sua morte [Beppu et al. 2021], o que é um limitante para as pesquisas. Neste sentido, foram usadas somente informações públicas na Web para montagem dos memoriais e testado com dados de um usuário vivo, que consentiu com o experimento.

Durante o desenvolvimento deste trabalho, foi identificadas limitações, como avaliação restrita a dois perfis (pesquisador + professor), o que reduz a generalização estatística; dependência de fontes públicas, dados privados ou protegidos não são incorporados; e possibilidade de imprecisões em categorias limítrofes (p. ex., hobbies × vida pessoal) que ainda exigem revisão humana. Além disso, a ausência de conteúdo multimídia estruturado (descrições automáticas de imagens/vídeos) e o uso de um modelo relativamente compacto (*Gemma 3-1B*) impõem limites à riqueza das narrativas e à personalização estilística.

Para trabalhos futuros, recomendamos (i) testar modelos maiores ou multilingues em busca de resumos mais precisos e dialetos diversificados; (ii) integrar visão computacional para gerar descrições automáticas de fotos e vídeos; (iii) conduzir estudos de usabilidade com familiares enlutados para refinar interface e tom comunicativo; e (iv) conectar a ferramentas de arquivo web (*Archive.org*, *Arquivo.pt*) para importar históricos completos, ampliando a cobertura factual. Essas extensões visam fortalecer a missão de salvaguardar legados digitais de forma escalável, humanizada e culturalmente sensível.

Agradecimentos

Os autores agradecem o apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Pró-Reitoria de Pesquisa (PROPeq-UFMT). Foram utilizadas ferramentas de IA para auxiliar na formatação e revisão de pontuação do texto, com todo o conteúdo revisado e validado pelos autores.

Referências bibliográficas

- Beppu, F., Maciel, C., & Viterbo, J. (2021). Contributions of the Brazilian Act for the Protection of Personal Data for treating Digital Legacy. *Journal on Interactive Systems*, 12, 112-124. Disponível em: <https://doi.org/10.5753/jis.2021.1654>. Acesso em: 20 ago. 2025.
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.
- Brown, M. A., Gruen, A., Maldoff, G., Messing, S., Sanderson, Z., & Zimmer, M. (2024). *Web Scraping for Research: Legal, Ethical, Institutional, and Scientific*

- Considerations. arXiv preprint arXiv:2410.23432. Disponível em: <https://arxiv.org/abs/2410.23432>. Acesso em: 20 ago. 2025.
- Creswell, J. W., Creswell, J. D. (2018). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications, Los Angeles, CA, 5th edition.
- Dados Além da Vida (DAVI). (2025). DAVI | Dados Além da Vida. Disponível em: <https://lavi.ic.ufmt.br/davi/>. Acesso em: 20 ago. 2025.
- Gemma Team. (2025). Gemma 3 Technical Report. arXiv preprint arXiv:2503.19786. Disponível em: <https://arxiv.org/abs/2503.19786>. Acesso em: 20 ago. 2025.
- Gil, A. C. (2019). *Como elaborar projetos de pesquisa*. 6. ed. São Paulo: Atlas.
- Lopes, A. D., Maciel, M., & Pereira, V. C. (2014). Recomendações para o design de memórias digitais na web social. In *Proceedings of the 13th Brazilian Symposium on Human Factors in Computing Systems (IHC '14)*. Sociedade Brasileira de Computação, 275–284.
- McKinney, W. (2010). Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference* (pp. 51-56).
- Ni, J., Li, J., & McAuley, J. (2021). Sentence-T5: Scalable sentence encoders from pre-trained text-to-text models. arXiv preprint arXiv:2108.08877. Disponível em: <https://arxiv.org/abs/2108.08877>. Acesso em: 20 ago. 2025.
- Öhman, C. J., & Watson, D. (2019). Are the dead taking over Facebook? A Big Data approach to the future of death online. *Big Data & Society*, 6(1), 2053951719842540. Disponível em: <https://doi.org/10.1177/2053951719842540>. Acesso em: 20 ago. 2025.
- Pereira, R.; et al. (2024). GranDIHC-BR: Grand Research Challenges in Human-Computer Interaction in Brazil for 2025–2035. *IHC 2024 – Anais Estendidos*. Disponível em: <https://dl.acm.org/doi/10.1145/3702038.3702061>. Acesso em: 20 ago. 2025.
- Reitz, K. (2025). Requests: HTTP for Humans. Disponível em: <https://requests.readthedocs.io/en/master/>. Acesso em: 20 ago. 2025.
- Richardson, L. (2025). Beautiful Soup. Disponível em: <https://www.crummy.com/software/BeautifulSoup/>. Acesso em: 20 ago. 2025.
- Selenium Developers. (2025). Selenium WebDriver. Disponível em: <https://www.selenium.dev/>. Acesso em: 20 ago. 2025.
- Sommerville, I. (2015). *Software Engineering*. 10th Edition, Pearson Education Limited, Boston.
- Trevisan, D., Maciel, C., & Bim, S. A. (2021). Educação, morte e tecnologias - experiência no ensino de avaliação em IHC. In *Anais Estendidos do XX Simpósio Brasileiro de Fatores Humanos em Sistemas Computacionais (IHC)*, 56-63. Porto Alegre: SBC. Disponível em:

https://sol.sbc.org.br/index.php/ihc_estendido/article/view/19590. Acesso em: 20 ago. 2025.

Ueda, G., Monteiro, L. F. F., Maciel, C., & Pereira, V. C. (2022). Digital memorials: classifications and design recommendations. *Journal on Interactive Systems*, 13(1), 335–349. Disponível em: <https://journals-sol.sbc.org.br/index.php/jis/article/view/2567>. Acesso em: 20 ago. 2025.

Wolf, T., et al. (2019). HuggingFace’s Transformers: State-of-the-art Natural Language Processing. arXiv 1910.03771. Disponível em: <https://arxiv.org/abs/1910.03771>. Acesso em: 20 ago. 2025.