Título em português: PROJETO PERPETUA: USO DE BIBLIOTECAS PYTHON EM FLUXO DE TRABALHO PARA PRESERVAÇÃO DE ACERVOS ARQUIVÍSTICOS LEGADOS

Título em Inglês: PERPETUA PROJECT: USING PYTHON LIBRARIES IN A WORKFLOW FOR PRESERVING LEGACY ARCHIVAL COLLECTIONS

Resumo: O projeto visa promover a preservação digital de conjuntos documentais legados em uma instituição arquivística que enfrenta desafios comuns a muitas outras entidades de pequeno e médio porte. O cenário típico envolve acervos com documentos em papel e digitalizações realizadas sem o devido controle sobre a correspondência entre os arquivos digitais e suas informações descritivas. Esse problema gera duplicações e versões redundantes dos documentos digitais, dificultando a recuperação da informação. A solução proposta é reaproveitar digitalizações existentes, mesmo sem padronização ou organização prévia, evitando a re-digitalização e seus altos custos. A metodologia adotada é adaptável e semi-automatizada, utilizando bibliotecas *open source,* permitindo que as instituições estruturem seus acervos digitais sem comprometer recursos. O modelo também contribui para a padronização de práticas em um fluxo de trabalho de gestão arquivística e preservação digital, facilitando a colaboração entre instituições e a adoção de estratégias acessíveis e eficazes, seguindo a política de preservação da instituição, de acordo com padrões nacionais e internacionais. Os resultados, até o momento, mostram a viabilidade da abordagem cuja principal vantagem é o reaproveitamento das digitalizações anteriores, sem a necessidade de refazer os processos, mantendo registros padronizados e melhorando a rastreabilidade e controle arquivístico. A conclusão reafirma a importância de políticas de preservação e de padrões desde a produção do documento, destacando que é possível aplicar critérios técnicos que assegurem a integridade, autenticidade e acessibilidade dos acervos digitais a longo prazo, sem a necessidade de descartes ou reprocessamentos dispendiosos.

Palavras-chave: preservação digital, gestão arquivística, automação de processos, biblioteca Python, pacote de informação.

Abstract: The project aims to promote the digital preservation of legacy document collections within an archival institution facing challenges common to many small and medium-sized entities. The typical scenario involves collections comprising both paper records and previously digitized materials, often lacking adequate control over the correspondence between digital files and their descriptive metadata. This leads to duplication and redundant versions of digital records, hindering effective information retrieval.The proposed solution centers on reusing existing digitizations, regardless of their lack of standardization or prior organization, thus avoiding costly re-digitization processes. The adopted methodology is adaptable and semi-automated, using open-source libraries, allowing institutions to structure and manage their digital collections without placing a heavy burden on their resources.Moreover, the model contributes to the standardization of archival workflows for digital preservation, supporting institutional collaboration and the adoption of accessible, cost-effective strategies. It aligns with institutional preservation policies and complies with both national and international standards.Preliminary results confirm the feasibility of the approach, with the main advantage being the reuse of prior digitizations without the need to repeat processing steps. This supports the creation of standardized records, enhances traceability, and strengthens archival control. The conclusion underscores the critical role of preservation policies and standards beginning at the point of record creation. It emphasizes that it is indeed possible to apply technical criteria that ensure the long-term integrity, authenticity, and accessibility of digital collections without the need for costly disposal or reprocessing efforts.

keywords: digital preservation, record management, task automation. Python library, information package

Temática: Curadoria Digital, Arquivística e Preservação Digital

Introdução

Esse projeto foi desenvolvido em uma instituição arquivística com vistas a promover a preservação digital de conjuntos documentais legados que fazem parte de seu acervo. No entanto, sua concepção considerou um cenário amplamente compartilhado por muitas outras instituições do mesmo gênero, especialmente de pequeno e médio porte. Tal cenário se caracteriza pela predominância de conjuntos documentais com suporte em papel e parcial ou integralmente digitalizados, mas cujas digitalizações, as quais chamaremos de representantes digitais, foram geradas sem estrito controle de correspondência entre os representantes e suas informações descritivas.

Sem o devido controle, ou seja, sem que os representantes digitais sejam corretamente nomeados, descritos e indexados, um mesmo documento pode ser digitalizado várias vezes com nomes e formatos distintos. Inevitavelmente, essa prática resulta em duplicações e versões redundantes que, longe de constituírem um sistema de backup, apenas oneram sobremaneira o espaço de armazenamento disponível e inviabilizam a recuperação da informação, agravando a desorganização e prejudicando a eficiência administrativa.

Nesse contexto, a solução mais simples costuma ser descartar os arquivos digitais existentes e refazer as digitalizações após a definição dos padrões a serem seguidos e do processo de organização e classificação documental, uma vez que digitalizações geradas fora de uma estrutura de classificação não têm vínculo com processos, séries ou fundos e pode ser inútil do ponto de vista histórico, administrativo ou jurídico ainda que seu conteúdo seja legível. No entanto, o custo dessa re-digitalização pode superar os recursos disponíveis da entidade custodiadora, ainda mais quando parte significativa das digitalizações já passou por processamento técnico para melhorar a qualidade visual das informações, tornando a medida inviável do ponto de vista orçamentário.

Por oferecer um modelo de ação adaptável e semi-automatizado que pode ser aplicado por outras instituições que enfrentam desafios semelhantes, permitindo o reaproveitamento de acervos digitais existentes e a estruturação de práticas voltadas à gestão arquivística, entendemos que é do interesse coletivo difundir essa experiência, de modo a contribuir com a consolidação de estratégias acessíveis e eficazes para a salvaguarda do patrimônio documental digital. Ao compartilhar os resultados e as metodologias adotadas, espera-se fortalecer a cooperação entre instituições e fomentar a adoção de práticas sustentáveis e escaláveis no campo da arquivística e da preservação digital.

Metodologia

Diante da necessidade de garantir que os representantes digitais sejam preservados em pacotes de informação que contenham todos os elementos essenciais à sua interpretação e reutilização no longo prazo, adotamos o formato BagIt como padrão de empacotamento. Essa decisão foi motivada pela simplicidade estrutural do BagIt, sua robustez na verificação de integridade e sua ampla adoção por instituições voltadas à preservação digital, o que favorece

a interoperabilidade e a sustentabilidade dos acervos ao longo do tempo. Para assegurar que os pacotes sejam compreensíveis, foi necessário modelar requisitos mínimos, tais como padrões de metadados, regras de nomenclatura, formatos de preservação e parâmetros técnicos, como a resolução de digitalização. A partir dessa modelagem, e com base nas diretrizes do modelo *Open Archival Information System* - OAIS (ISO, 14721:2025), das Resoluções nº 48 (BRASIL, 2021) e nº 51 do Conselho Nacional de Arquivos - CONARQ (BRASIL, 2023) e do *Digital Preservation Toolkit* (CHIN, 2011), a implementação da metodologia ocorreu em duas frentes distintas, que resultaram em produtos independentes, mas concebidos para serem aplicados de forma complementar.

De um lado, realizou-se o mapeamento dos processos de tratamento arquivístico da instituição, desde o recolhimento até a digitalização e o armazenamento dos representantes digitais. A partir desse levantamento, foi proposto e aperfeiçoado um **fluxo de trabalho** utilizando a notação BPMN (*Business Process Model and Notation*), com o objetivo de padronizar as práticas, reduzir erros no cumprimento dos requisitos e aumentar a eficiência operacional.

De outro lado, foi desenvolvida a **biblioteca Python Perpetua**, baseada exclusivamente em dependências *open source*, voltada à automação de tarefas repetitivas sem gerar custos adicionais para a instituição. Dentre suas funcionalidades principais, destacam-se:

Inventário de representantes digitais: geração de um inventário recursivo dos objetos digitais, contendo nome, hash, tamanho, formato e localização. Esse inventário serve de base para a avaliação e adequação dos procedimentos arquivísticos adotados.

Padronização de nomenclatura: após a seleção do conjunto documental a ser preservado em meio digital, um módulo da biblioteca extrai expressões regulares e, com base nas regras definidas para a nova nomenclatura, gera uma lista em duas vias com as denominações recomendadas. Uma das vias é submetida à análise do arquivista, que poderá incluir (em caso de documentos ainda não digitalizados) ou excluir linhas (em caso de duplicações). Um script subsequente realiza a renomeação dos arquivos com base na lista revisada.

Normalização de formatos: script responsável por converter os arquivos existentes para os formatos de preservação definidos, assegurando a geração de objetos derivados 1:1 (um objeto digital para cada documento arquivístico), mesmo em casos em que o documento original tenha sido digitalizado em partes.

Verificação de conformidade de estrutura e informações: um módulo que avalia se as regras estabelecidas para o preenchimento de metadados descritivos foram corretamente seguidas e identifica possíveis inconsistências entre os objetos digitais e seus respectivos metadados.

Montagem de pacotes: Geração de pacotes estruturados conforme os requisitos da plataforma de preservação, de modo que possam ser convertidos ao padrão *BagIt*. Quando necessário, os pacotes são automaticamente divididos em subconjuntos menores para facilitar a ingestão e o armazenamento.

Em todas as etapas, inclusive nas não descritas aqui, a movimentação dos objetos digitais é realizada de forma segura, com verificação de checksums após cada operação. Todos os resultados e saídas do processo são preservados em formatos estruturados como JSON, PDF ou CSV, conforme a natureza e a conveniência dos dados representados.

Resultados

Com base na experiência da instituição, os resultados obtidos demonstram a viabilidade e os benefícios da metodologia adotada. Um fundo já foi completamente processado, totalizando cerca de 18.400 objetos digitais, entre matrizes (1:n) e derivadas (1:1), e outro fundo encontra-se em fase de processamento. Entre as principais vantagens observadas, destaca-se a possibilidade de reaproveitar digitalizações previamente realizadas, mesmo que tenham sido produzidas sem critérios de organização ou padronização, e a criação de registros padronizados em todas as etapas de preparação dos pacotes de informação. Tais registros, juntamente com as informações desestruturadas produzidas acerca da documentação muito antes da implementação da metodologia, foram preservadas integralmente nos pacotes de informação correspondentes promovendo maior consistência, rastreabilidade e controle arquivístico, sem que fosse necessário descartar ou repetir os processos de digitalização.

Conclusão

A adoção dessa metodologia não prescinde da definição de uma política de preservação e muito menos da implementação de padrões desde o momento de produção do documento, pelo contrário, reafirma sua necessidade em vista do trabalho necessário para reajustar o que não precisaria de ajuste se feito corretamente. Todavia, permite que entidades custodiadoras que estão no processo de revisão das práticas para adequação de acervos aos requisitos da preservação digital encontrem um caminho viável para estruturar seus acervos digitais, mesmo diante de um cenário inicial desorganizado. Ao reconhecer os erros do passado e aplicar critérios técnicos a partir de uma política bem definida, essas instituições podem avançar rumo à integridade, autenticidade e acessibilidade de longo prazo dos documentos sob sua guarda.

Referência bibliográfica

BRASIL. Conselho Nacional de Arquivos (CONARQ). **Resolução Conarq nº48, de 10 de novembro de 2021**. Estabelece diretrizes e orientações aos órgãos e entidades integrantes do Sistema Nacional de Arquivos quanto aos procedimentos técnicos a serem observados no processo de digitalização de documentos públicos ou privados. Disponível em: <https://www.gov.br/conarq/pt-br/centrais-de-conteudo/publicacoes/Diretrizes_digitalizacao__2021.pdf>. Acesso em: 03 jun. 2024.

BRASIL. Conselho Nacional de Arquivos (CONARQ). **Resolução nº 51, de 25 de agosto de 2023**. Dispõe sobre as "Diretrizes para a Implementação de Repositórios Arquivísticos Digitais Confiáveis", Versão 2. Disponível em: <https://www.gov.br/conarq/pt-br/legislacao-arquivistica/resolucoes-do-conarq/resolucao-conarq-no-51-de-25-de-agosto-de-2023>. Acesso em: 03 jun. 2024.

CHIN. Canadian Heritage Information Network. **Digital Preservation Toolkit**. 2021-11-25. Disponível em: <https://www.canada.ca/en/heritage-information-network/services/digital-preservation/toolkit.html>. Acesso em: 23 de maio de 2025.

ISO. **Space data system practices — Reference model for an open archival information system (OAIS)**. ISO 14721:2025(en). Geneva: International Organization for Standardization, 2025.