Tecnologias da web e inteligência artificial
novas fronteiras para coleções digitais

Web technologies and artificial intelligence: new frontiers for digital collections / Tecnologías web e inteligencia artificial: nuevas fronteras para las colecciones digitales

Juliana Marques
Doutora em Sociologia pela Universidade do Estado do Rio de Janeiro (Uerj). Professora do Programa de Pós-Graduação em História, Política e Bens Culturais da Fundação Getúlio Vargas (FGV), Brasil.
juliana.marques@fgv.br

Suemi Higuchi
Doutora em Estudos da Linguagem pela Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio). Pesquisadora e professora da Escola de Ciências Sociais da Fundação Getúlio Vargas (FGV), Brasil.
suemi.higuchi@fgv.br

RESUMO

A digitalização de coleções históricas e culturais tem precipitado um processo de profunda transformação, conhecido como "datificação" das fontes primárias. Esta evolução reconfigura a natureza e o potencial dessas coleções, gerando dados acessíveis e reutilizáveis, ampliando as possibilidades de pesquisa e democratizando o acesso ao patrimônio. Neste artigo, exploramos essa transformação e suas implicações para o futuro de instituições de pesquisa e de guarda de acervos histórico-culturais.

Palavras-chave: dados abertos; web semântica; inteligência artificial; humanidades digitais.

ABSTRACT

The digitization of historical and cultural collections has precipitated a process of profound transformation, known as the "datafication" of primary sources. This evolution reconfigures the nature and potential of these collections, generating accessible and reusable data, expanding research possibilities and democratizing access to heritage. In this article, we explore this transformation and its implications for the future of research institutions and the custody of historical and cultural collections.

Keywords: open data; semantic web; artificial intelligence; digital humanities.

RESUMEN

La digitalización de colecciones históricas y culturales ha precipitado un proceso de profunda transformación, conocido como "datificación" de las fuentes primarias. Esta evolución reconfigura la naturaleza y el potencial de estas colecciones, generando datos accesibles y reutilizables, ampliando las posibilidades de investigación y democratizando el acceso al patrimonio. En este artículo, exploramos esta transformación y sus implicaciones para el futuro de las instituciones de investigación y la custodia de las colecciones histórico-culturales.

Palabras clave: datos abiertos; web semántica; inteligencia artificial; humanidades digitales.

Introdução

A emergência da chamada web semântica traz implicações positivas para a democratização do acesso à informação, com impacto para a forma de atuação de arquivos, museus e bibliotecas. Quando a internet foi concebida,[1] a promessa era a de uma rede de comunicação global e descentralizada que democratizaria o acesso à informação e promoveria a colaboração em todo o mundo. Esse "todo o mundo", claro, seria quem estivesse conectado a ela.

Se, por um lado, é verdade que foram feitos grandes avanços na oferta de acesso generalizado à internet, com sete bilhões de pessoas conectadas ao redor do globo, sendo mais de 180 milhões de pessoas[2] no Brasil (TIC Domicílios, 2023), por outro, desigualdades sociais preexistentes são reforçadas e repaginadas como desigualdades sociodigitais. Além disso, com o surgimento de grandes empresas de tecnologia, não conseguimos evitar monopólios e pontos centrais de controle sobre a operação desse novo e dinâmico mundo digital.

Desde as bibliotecas da antiguidade e monásticas até as instituições culturais modernas, esses espaços de guarda da informação e de produção do conhecimento também desempenharam papéis contraditórios. De um lado, dispositivos de poder-saber que refletiam e reforçavam estruturas de poder existentes, determinando quem poderia acessar quais informações e sob que circunstâncias, e, de outro, heterotopias – espaços disruptivos e transformadores que dão lugar a ideias e ações não-hegemônicas (Foucault, 2013).

A partir dos movimentos a favor da educação pública e da democratização do conhecimento no século XIX, observou-se uma profunda transformação no éthos e nas missões dessas instituições. O surgimento de bibliotecas públicas na Inglaterra, França e nos Estados Unidos, por exemplo, marcou uma mudança significativa na concepção do acesso à informação como um direito civil. Paralelamente, museus passaram a repensar suas práticas curatoriais e educacionais, buscando engajar um público mais amplo e diverso. A revolução digital veio aprofundar e acelerar essas mudanças, junto a esforços para descolonizar as coleções (Jones, 2017; Liddington, 2002).

A web semântica ou web de dados, enquanto uma extensão da world wide web que já conhecemos, visa tornar o conteúdo on-line mais acessível e palatável para humanos e mais processável por máquinas, possibilitando buscas mais

1 O Computer History Museum apresenta uma linha do tempo detalhada sobre a história da internet, acessível em: https://www.computerhistory.org/internethistory.

2 Equivalente a 84% da população com dez anos ou mais.

precisas, porque também mais contextualizadas. Essa inovação tem potencial de promover transparência e acessibilidade, além da inclusão de diversas vozes e perspectivas. As instituições de guarda desempenham um papel fundamental nesse processo. Contudo, a integração de dados de diferentes fontes e a recuperação mais sofisticada de informação, baseadas em dados estruturados e interligados e no uso da inteligência artificial, apresentam desafios significativos de padronização e implementação. Ainda assim, essas tecnologias emergentes prometem revolucionar a maneira como utilizamos a web.

Diversas mudanças já foram introduzidas com a web 2.0[3] a partir dos anos 2000, quando aprimoramos a nossa comunicação com websites, blogs, redes sociais, e-mails, mensagens instantâneas e videoconferências. Plataformas de busca na web, como o Google, bibliotecas e repositórios digitais, além de bancos de dados e periódicos on-line de acesso aberto ampliaram e diversificaram a forma como construímos e obtemos conhecimento. A internet tem sido, de fato, um catalisador para a inovação, criando indústrias inteiras e transformando outras. A web fornece lócus e ferramentas para movimentos sociais e múltiplos ativismos políticos, dando voz, texto, imagem e interconexões a grupos invisibilizados por estruturas antigas de poder. Em que fronteira estamos agora, com a web 3.0?

Arquivos, museus e bibliotecas não são instituições em processo de obsolescência frente a dispositivos digitais e a web semântica.[4] Pelo contrário, elas continuam sendo fundamentais para a construção de memórias e narrativas coletivas sobre o passado e o presente e, cada vez mais, incluem perspectivas contra-hegemônicas, produzidas a partir de inovadoras práticas patrimoniais, com a participação direta e o engajamento de diferentes grupos sociais (Boita, Baptista, 2024; Brown et al., 2023; Knudsen et al., 2021). Além de preservar e disseminar registros que se transformam em patrimônio histórico e cultural, essas instituições são, então, ativas na criação e na interpretação de conhecimento, oferecendo contexto e significado às informações que acumulam e ajudam a disseminar.

3 A primeira fase da web era principalmente estática e composta por páginas com conteúdo fixo, nas quais os usuários eram apenas consumidores passivos de informação. A web 2.0 inaugurou uma internet mais interativa e dinâmica, permitindo que os usuários também criassem e compartilhassem conteúdo através de blogs, redes sociais e plataformas de colaboração.

4 Nela, através de padrões e tecnologias específicas, como o RDF (Resource Description Framework), adicionamos significado e contexto às informações que circulam, melhorando a interoperabilidade e a reutilização de dados entre diferentes sistemas. Como resultado, os computadores entendem e processam a informação de maneira mais inteligente e o usuário consegue fazer buscas de informações mais precisas e eficientes.

Como sociedade global mais interconectada, dispomos de uma oportunidade sem precedentes. Por um lado, podemos compartilhar experiências e reflexões que promovem a inclusão e a democratização com uma facilidade nunca vista. Por outro, temos a capacidade de integrar recursos informacionais de diversas fontes e de enriquecer e interconectar dados de formas mais inteligentes. Nesse universo on-line cada vez mais diversificado, urge a necessidade de desenvolvermos formas avançadas de análise, curadoria e apresentação da informação.

Neste artigo compartilhamos insights de nossa jornada de investigação para a implementação de um sistema-piloto de gestão e exposição de coleções digitais para o acervo do Centro de Pesquisa e Documentação de História Contemporânea do Brasil (Cpdoc) da Fundação Getúlio Vargas (FGV), guiadas pelos princípios de dados abertos. É o resultado de nossa experiência prática e de leituras, análises de experiências de outras instituições, estudos de manuais e melhores práticas, além de documentações de sistemas de gestão de acervos e padrões existentes. Representa um passo importante em nossa busca pela solução ideal, oferecendo uma base sólida para futuras decisões e implementações.

Na primeira parte, resumimos o impacto mais amplo da digitalização nas instituições culturais que abrigam acervos histórico-culturais. Nas três seções seguintes, abordamos a abertura de dados de coleções digitais, dedicamos uma seção à revisão técnica sobre padrões de metadados, ontologias e protocolos de comunicação, e explicamos jargões para descomplicar os meandros da interoperabilidade, apresentando exemplos sempre que possível. Em seguida, passamos às duas últimas seções, nas quais discutimos experiências no Cpdoc, com exemplos de estruturação e integração de dados na web e de novas abordagens com o uso de inteligência artificial. Na conclusão, sintetizamos as principais lições aprendidas e refletimos sobre as profundas transformações pelas quais o trabalho em instituições de guarda e memória está passando.

O impacto da digitalização nas instituições de guarda

Arquivos, museus e bibliotecas agora precisam gerenciar tanto coleções físicas quanto digitais, exigindo o desenvolvimento de novas habilidades e infraestruturas para lidar com formatos digitais, metadados, preservação digital e acesso on-line. A inexorabilidade do processo de digitalização da vida abarca, enfim, o sensível trabalho de arquivistas, historiadores e outros profissionais que lidam com esses registros que constituem a base empírica do nosso conhecimento e das nossas interpretações, permitindo-nos construir e relativizar a compreensão da nossa história e identidade.

Esse é um processo que se avoluma desde os anos 2000, quando, pouco a pouco, instituições de guarda passam a digitalizar seus acervos. No Brasil de 2022, 84% dos arquivos e 68% dos museus haviam digitalizado parte de seu acervo. Quando apuramos o foco para aquelas instituições que disponibilizaram esse material digitalizado on-line, temos 64% dos arquivos e 35% dos museus (TIC Cultura, 2023).

Disponibilizar um acervo on-line é importante porque significa expandir seu alcance a um público global e diversificado.[5] E não é uma tarefa trivial, pois requer um sistema de gestão de objetos digitais, de preservação digital e de recuperação da informação, além de uma interface para o usuário fazer suas consultas e acessar o material. Uma vez vencidos esses desafios, é comum que a próxima etapa de integração de tecnologias digitais passe pela disseminação das coleções nas redes sociais. É crescente o uso das redes sociais, por parte dos equipamentos culturais brasileiros, conforme mostra a série histórica do *survey* TIC Cultura (2023). Em 2022, uma parcela significativa dos participantes da pesquisa – 58% dos arquivos e 56% dos museus – relatou utilizar as redes sociais para divulgar seus acervos, projetos ou serviços.

Do ponto de vista técnico, essa adoção representa um avanço de baixa complexidade. Ainda que o trabalho de curadoria seja importante, demandando conhecimento especializado e familiaridade com as coleções, a web 2.0 trouxe consigo uma gama de ferramentas intuitivas que facilitam esse processo. O Instagram, por exemplo, tornou-se uma plataforma popular para compartilhamento de conteúdo visual, enquanto aplicativos de design gráfico como o Canva permitem que mesmo usuários sem experiência criem materiais de divulgação atraentes para suas postagens. A relevância de potencializar esses espaços como estratégias importantes para estabelecer a presença digital dessas instituições cresce ainda mais ao considerar que muitas delas carecem de sites institucionais e dispõem apenas das redes sociais para disseminar seus acervos (Martins, 2020, p. 103).

Para além disso, no entanto, o que podemos fazer, como usar as tecnologias digitais a nosso favor, em prol da missão dessas instituições? Que outras inovações tecnológicas podem ser desenvolvidas ou adaptadas para enriquecer a experiência do público, preservar o patrimônio cultural de forma mais eficaz e ampliar o alcance desses espaços?

5 É crucial integrar as dimensões on-line e off-line, garantindo que a digitalização amplie o acesso sem comprometer a preservação física dos documentos. A preservação dos originais é essencial, pois eles carregam valor histórico, material e simbólico que não pode ser totalmente capturado ou garantido em formato digital.

O que nem sempre está evidente desde o princípio é que a digitalização de fontes primárias – sejam elas fotografias, arquivos de áudio, de texto, de vídeo ou mesmo objetos que passem a ter o seu modelo 3D – permite a criação de novas camadas de informação sobre cada um desses itens, abrindo múltiplas possibilidades de pesquisa, educação e engajamento público com o patrimônio cultural. Para além dos metadados descritivos e de proveniência mais tradicionais, os objetos digitalizados ou nato-digitais incorporam uma gama mais ampla de informações. Eles trazem metadados técnicos, por exemplo, sobre o processo de digitalização e de preservação, e podem conter transcrições textuais, dados de geolocalização, relação com outros objetos digitais e links externos, além de variantes digitais com diferentes representações de si.

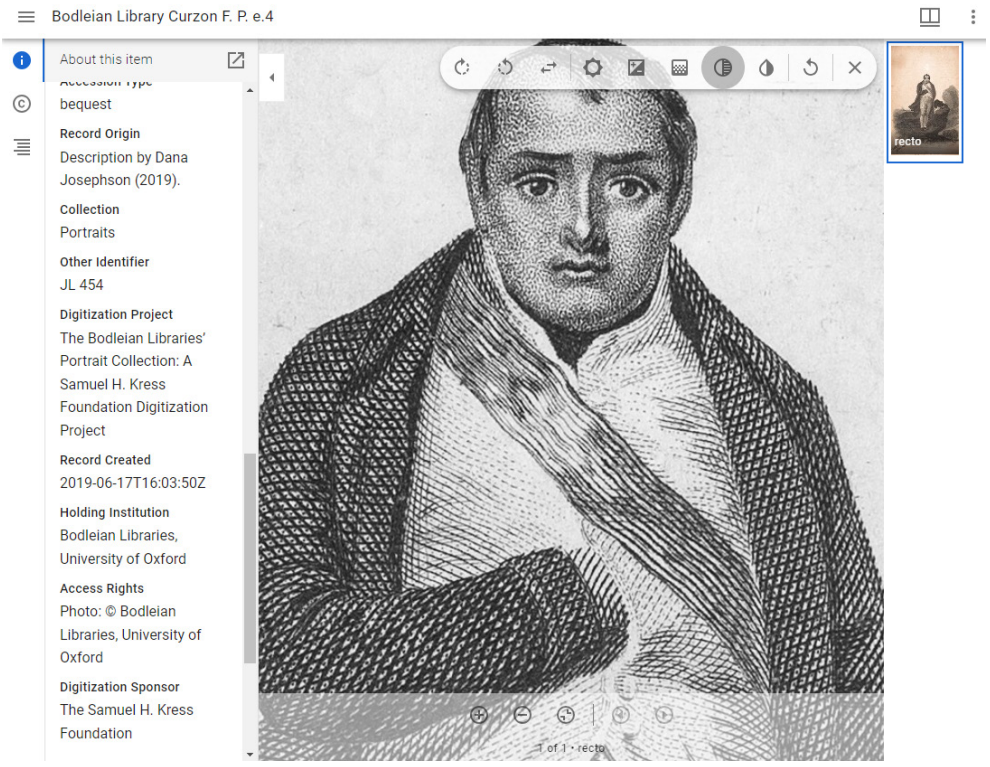

Figura 1 – Interface de visualização com metadados e diferentes representações digitais de um único objeto. Retrato de Napoleão Bonaparte em gravura, exibido no visualizador Mirador/IIIF. A interface mostra, no painel esquerdo, os metadados detalhados, incluindo licença de uso (access rights) e o patrocinador da digitalização (digitization sponsor). À direita, há duas visualizações do objeto digital, sendo uma miniatura em cor e a imagem ampliada em escala de cinza, que nos permite atentar para mais detalhes do desenho. Fonte: Oxford, Bodleian Library Curzon F. P. e.4. Disponível em: https://digital.bodleian.ox.ac.uk/objects/aaf4832b-c31d-4bbc-a2f3-278e553f0b74

Outras possibilidades incluem, ainda, a criação automática de metadados descritivos a partir do uso de inteligência artificial (IA) e a inclusão de anotações e *tags* do público usuário, seja ele especializado ou não. Ao integrar parte dos documentos textuais do Cpdoc na Infraestrutura Rossio,[6] como um dos objetivos de nosso projeto[7] de iniciação científica (IC), os títulos dos manuscritos, inexistentes na catalogação original, foram criados com inteligência artificial generativa e validados por nós. Na Figura 2, a interface Rossio exibe a lista de manuscritos do Cpdoc, com o detalhamento de um manuscrito selecionado. É possível visualizar a descrição original do item, um metadado da base de dados, juntamente com o título gerado pelo modelo de linguagem. Este é um dentre muitos usos possíveis de IA para o enriquecimento de informações e para a integração de acervos de diferentes instituições.

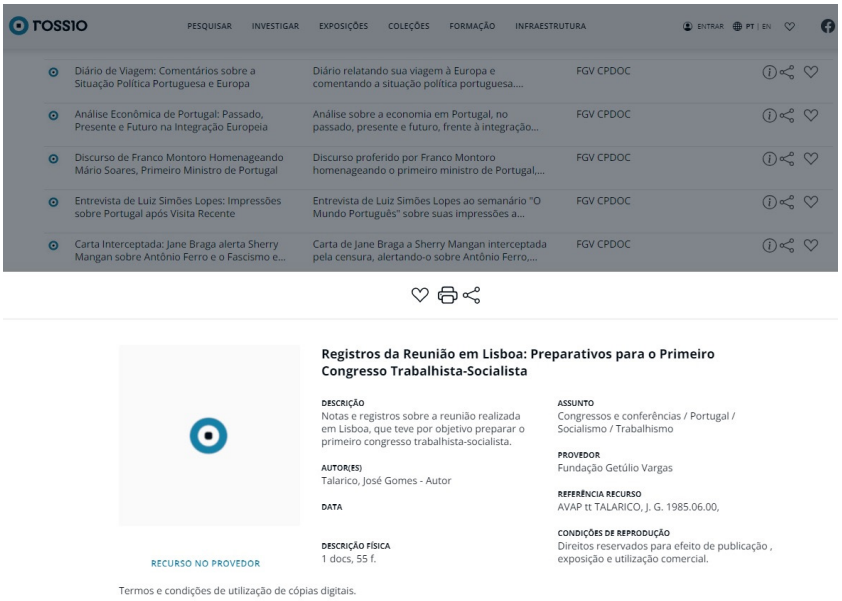

Figura 2 – Interface da infraestrutura Rossio com documentos do Cpdoc. Ao fundo, escurecida, a lista de documentos do Cpdoc. Abaixo, em destaque, são exibidos os metadados do item selecionado. Fonte: infraestrutura Rossio. Disponível em: https://rossio.pt/front/home

6 Um consórcio de instituições culturais que oferece fontes e recursos únicos sobre a diversidade da história, sociedade e patrimônio cultural de língua portuguesa. Disponível em: https://rossio.pt/.

7 "Tecnologias digitais e dados abertos para coleções históricas e de patrimônio cultural", com recursos da FGV, bolsa de IC do CNPq, bolsa de Iniciação Tecnológica da Faperj e com a participação de estudantes voluntários. O *script* de criação automatizada de títulos foi escrito pelo aluno Makalister Andrade.

Com essas novas camadas de informação digital, um mundo de possibilidades se abre. Há, no entanto, muitos passos a seguir no caminho da estruturação dos dados e na abertura desses acervos para amplo e diversificado uso na web. A "datificação" é um fenômeno da atualidade, é um desdobramento da aplicação de várias tecnologias digitais e métodos computacionais ao universo dos acervos. O aportuguesamento de *datafication* se refere, nesse contexto, à transformação de material de arquivo em dados digitais estruturados, que podem ser manipulados, analisados e empregados para diversos propósitos, utilizando o computador.

A abertura de dados de coleções digitais

O conceito de dados abertos, conforme definido pela Open Knowledge Foundation, refere-se a dados que podem ser livremente usados, reutilizados e redistribuídos por qualquer pessoa (Open Knowledge Foundation, 2012). Em uma definição mais elaborada, a cartilha da política de dados abertos do governo federal afirma que esses dados devem ser "acessíveis ao público, representados em meio digital, estruturados em formato aberto, processáveis por máquina, referenciados na internet e disponibilizados sob licença aberta que permita sua livre utilização, consumo ou cruzamento, limitando-se a creditar a autoria ou a fonte" (São Paulo, 2022, p. 49). A combinação desses atributos visa promover ao mesmo tempo a transparência e a inovação, permitindo que a comunidade de usuários desenvolva novas análises, aplicações e soluções baseadas nessas informações.

Há quase vinte anos, Tim Berners-Lee propôs uma web de dados abertos interligados (LOD, do inglês, Linked Open Data), introduzindo o esquema das cinco estrelas como critério de qualidade para o grau de abertura dos dados (Marcondes, 2021, p. 36). Este esquema progressivo abrange desde a simples disponibilização de recursos na web sob licença aberta até a sua finalidade maior, que é a conexão de dados com outras fontes para prover contexto. Parte integrante da visão da web semântica, as tecnologias LOD representam uma evolução na forma de publicar conteúdos, permitindo que os usuários descubram informações adicionais, compreendam melhor o contexto dos dados e combinem diferentes *datasets* de forma mais eficiente.

Para ilustrar, imagine uma coleção digital contendo documentos sobre a história indígena no Brasil. Neste nível de abertura, cada item da coleção possuiria um identificador único (URI) na web, permitindo uma interligação rica e contextualizada. Os metadados, como data, autor e local, seriam vinculados a outros conjuntos de dados relevantes. Por exemplo, nomes de povos indígenas poderiam ser conectados a bases de dados etnográficas, oferecendo informações

adicionais sobre sua cultura e localização. Locais mencionados nos documentos seriam linkados a bases geográficas, facilitando visualizações em mapas, enquanto datas poderiam ser associadas a linhas do tempo históricas para contextualizar os eventos. Nomes de autores ou pessoas citadas nos documentos poderiam ser vinculados a bases biográficas ou genealógicas, e termos específicos em línguas indígenas poderiam ser ligados a dicionários ou glossários on-line. Além disso, referências a outros documentos na coleção ou em outras instituições seriam diretamente vinculadas.

Essa estrutura interconectada proporcionaria uma experiência de pesquisa muito mais rica e contextualizada. Um pesquisador analisando um documento sobre, por exemplo, a demarcação de terras de um povo indígena na década de 1970, poderia facilmente acessar informações atualizadas sobre aquele povo, visualizar a localização geográfica precisa, entender o contexto histórico do período, explorar documentos relacionados em outras instituições e compreender termos em línguas indígenas usados no documento. Essa interconexão profunda de dados não apenas simplifica a pesquisa, mas também abre portas para descobertas inesperadas e uma compreensão mais abrangente e nuançada do material histórico, potencializando significativamente o valor e a utilidade da coleção digital.

Nesse horizonte de possibilidades, várias dimensões importantes devem ser consideradas, abrangendo aspectos técnicos, legais, éticos e práticos para a disponibilização pública desses acervos. Na dimensão técnica, é essencial uma abordagem padronizada e abrangente. O uso de formatos abertos e interoperáveis,[8] como o XML, CSV e JSON, garante que os documentos digitais possam ser acessados e interpretados por diferentes sistemas e plataformas. A infraestrutura tecnológica deve suportar o armazenamento seguro dos dados e proporcionar acesso eficiente e escalável. A adoção de padrões de metadados e ontologias bem estabelecidos é crucial para organizar e descrever os acervos de forma consistente, facilitando a descoberta e o entendimento dos conteúdos.

8 Formatos de dados abertos, como XML (eXtensible Markup Language), CSV (Comma-Separated Values) e JSON (JavaScript Object Notation), são projetados para serem independentes de softwares proprietários, sendo usados para armazenar e transferir dados de maneira estruturada e eficiente. Esses formatos facilitam a troca de informações, garantindo compatibilidade, transparência e preservação a longo prazo, ao contrário do que ocorre com os formatos proprietários, como o DOCX (Microsoft Word), ou os formatos encapsulados, como o PDF (Adobe), que podem limitar a acessibilidade e o intercâmbio de dados entre diferentes plataformas e tecnologias.

Implementar protocolos de comunicação[9] e APIs[10] bem documentadas permite a integração fluida com outros sistemas e o desenvolvimento de aplicações por terceiros. A proveniência e rastreabilidade dos dados são igualmente importantes para manter a confiabilidade e o valor histórico dos acervos. Finalmente, estratégias de preservação digital a longo prazo são essenciais para assegurar a integridade e acessibilidade futura dos acervos, considerando a obsolescência tecnológica e a degradação de mídias digitais.

No contexto legal e ético, a abordagem deve ser cuidadosa e multifacetada. A adoção de licenças de uso adequadas, como as Creative Commons,[11] pode facilitar o acesso e a reutilização responsável dos materiais. É imprescindível a conformidade com leis de proteção de dados, como a Lei Geral de Proteção de Dados (LGPD) no Brasil, assegurando o tratamento ético de informações pessoais. Deve-se ter especial atenção às informações sensíveis ou potencialmente prejudiciais, estabelecendo protocolos para sua identificação e manejo apropriado. O respeito às culturas e comunidades representadas nos acervos é fundamental, envolvendo diálogo e consentimento quando apropriado. Assim, as instituições devem buscar um equilíbrio delicado entre a abertura de dados e a proteção de informações sensíveis, considerando o interesse público e as particularidades culturais em cada decisão de disponibilização.

A cultura do compartilhamento aberto ensejou iniciativas importantes no contexto das Glams (galerias, bibliotecas, arquivos e museus), como a parceria Glam Wiki (Monteiro, 2021). A partir dela, as instituições podem disponibilizar suas coleções ao público em formatos de dados abertos e licenciados na plataforma Wikimedia (incluindo Wikipédia, Wikimedia Commons e Wikidata), permitindo que o conteúdo seja acessado, reutilizado e remixado globalmente. Essa colaboração oferece benefícios mútuos: a comunidade Wikimedia se enriquece com a adição de conteúdos verificados e de alta qualidade e as instituições podem utilizar as ferramentas e o apoio da comunidade Wiki para catalogar e

9 Na internet, esses protocolos definem como os dados são transmitidos e recebidos entre diferentes sistemas ou dispositivos. Um dos protocolos mais conhecidos é o HTTP (Hypertext Transfer Protocol), que é a base para nossa navegação na web. Há diversos protocolos específicos para acervos e coleções digitais, como o OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) e o IIIF (International Image Interoperability Framework).

10 No contexto de arquivos e museus, uma API (Interface de Programação de Aplicações) é uma ferramenta que pode ser utilizada para integrar bases de dados e permitir que informações, como metadados, sejam acessadas e compartilhadas entre plataformas de forma automatizada e estruturada.

11 As licenças Creative Commons oferecem diversas opções para detentores de direitos autorais, como a CC BY-NC, que permite uso não comercial com atribuição. É considerada uma boa prática identificar claramente sob que licença os objetos ou coleções operam e quais itens estão em domínio público, facilitando assim o acesso e eliminando barreiras legais.

divulgar suas coleções. Isso não só reduz os custos associados à criação e manutenção de um portal próprio, mas também garante que os dados sejam apresentados em um formato compatível com as práticas de LOD, aumentando a visibilidade e o impacto cultural desses acervos em uma escala global.

Padrões de metadados, ontologias e protocolos de comunicação

No processo de abertura de dados, três elementos se destacam como pilares fundamentais: padrões de metadados, ontologias e protocolos de comunicação. Estes componentes, longe de serem meros detalhes técnicos, constituem a espinha dorsal de um ecossistema digital robusto e interconectado, crucial para instituições dedicadas à preservação e à disseminação do conhecimento contido em seus acervos digitais.

Metadados – ou metainformações – podem ser definidos como "dados que descrevem dados". Gilliland (2016) resume em três as características essenciais presentes em todos os objetos de informação, independentemente de seu formato: conteúdo, contexto e estrutura, todas elas refletidas através de metadados. O conteúdo se refere ao que o objeto contém ou de que ele trata; o contexto abrange as circunstâncias de sua criação e uso; e a estrutura se relaciona com as associações entre ou dentro dos objetos. Profissionais de bibliotecas, museus e centros de documentação utilizam metadados para organizar e facilitar o acesso a esses objetos, seguindo padrões desenvolvidos pela comunidade para garantir consistência e interoperabilidade. Segundo Boughida (2005), esses padrões podem ser divididos em quatro categorias principais: estrutura de dados, formato/intercâmbio técnico, valor de dados e conteúdo de dados, cada uma com exemplos específicos de aplicação.

Padrões de estrutura de dados, também chamados de esquemas ou conjuntos de elementos de metadados, são os contêineres de dados que compõem um registro ou objeto de informação. Alguns exemplos de padrões mais conhecidos nessa categoria são: o Marc (Machine-Readable Cataloging), o EAD (Encoded Archival Description), o Bibframe (Bibliographic Framework), o Dublin Core (Dublin Core Metadata Element Set) e o VRA Core (Works of Visual Resource). O conjunto de elementos de descrição, ou esquema de metadados, pode ser compartilhado para identificar recursos em outros sistemas de informação, e assim por diante.

Já os padrões de formato de dados/intercâmbio técnico são os padrões de metadados expressos em forma legível por máquina. Esse tipo de padrão é frequentemente uma manifestação de um padrão específico de estrutura de dados (conforme vimos acima), codificado ou marcado para processamento por

máquina. O profissional que lida com acervos e coleções digitais provavelmente já ouviu falar do RDF (Resource Description Framework), Marc21, MarcXML, EAD XML DTD, Mets, Lido XML, Dublin Core Simples XML, Dublin Core Qualificado XML ou VRA Core XML.

Os padrões de valor de dados são conhecidos pelos vocabulários controlados, tesauros e listas controladas.[12] São os termos, nomes e outros valores usados para preencher os padrões de estrutura de dados. Podemos mencionar aqui os cabeçalhos de assuntos da Biblioteca do Congresso Americano (Library of Congress Subject Headings), a lista de nomes de artistas (Union List of Artist Names), o tesauro de arte e arquitetura (Art & Architecture Thesaurus) e o tesauro de nomes geográficos (Thesaurus of Geographic Names), estes três últimos desenvolvidos pelo Getty Institute. No Brasil, museus, bibliotecas e arquivos também têm criado e compartilhado seus próprios vocabulários e tesauros, com o intuito de padronizar a terminologia e facilitar a indexação e recuperação de informações em acervos.

Por fim, padrões de conteúdo de dados dizem respeito às regras e aos códigos de catalogação. A Nobrade, por exemplo, é a norma que padroniza a descrição de documentos arquivísticos no Brasil. Desenvolvida pelo Conselho Nacional de Arquivos (Conarq) e inspirada na Isad(G) (International Standard for Archival Description) – norma internacional com alto grau de generalidade –, ela se constitui de elementos que orientam a descrição de documentos de arquivo, estruturando-os em níveis hierárquicos, desde o fundo até o item documental. Essa padronização garante a consistência e a interoperabilidade das informações, facilitando o acesso e a gestão de acervos entre as instituições. Outras normas importantes de descrição que vale a pena citar são: Isaar (CPF) (International Standard Archival Authority Record for Corporate Bodies, Persons and Families), Dacs (Describing Archives: A Content Standard), RAD (Rules for Archival Description) e RiC (Records in Contexts), cada qual com propósitos e escopos distintos.

Não existe um padrão único de metadados ou conjunto de padrões que seja adequado para descrever todos os tipos de coleções e materiais. Bibliotecas, museus e centros de documentação têm objetivos próprios de catalogação e descrição de objetos que integram seus acervos. A seleção do conjunto mais apropriado não só permitirá boas descrições de materiais de coleções diversas, como

12 A obra *Introdução aos vocabulários controlados*, publicada em 2016 pela Secretaria da Cultura do Estado de São Paulo, é uma importante referência sobre o assunto. Disponível em: https://www.sisemsp.org.br/wp-content/uploads/2023/03/vocabularios-controlados-digital.pdf. Acesso em: 10 ago. 2024.

também tornará possível mapear metadados criados a partir de necessidades específicas, promovendo assim o objetivo da interoperabilidade. Em muitos casos, especialmente com objetos complexos ou arquivos hierarquicamente estruturados, uma combinação de esquemas (por exemplo, Marc, Bibframe e/ou EAD no nível da coleção; Dublin Core, Mods, VRA Core e/ou Lido no nível do item) pode ser a melhor solução (Gilliland, 2016).

Diversos modelos de referência oferecem estruturas padronizadas e interoperáveis para a representação de informações no âmbito do patrimônio cultural. É o caso da EDM (Europeana Data Model), um modelo de dados que utiliza elementos de ontologias preexistentes e padrões de metadados, proporcionando um *framework* robusto para descrever e interconectar recursos culturais. A EDM não apenas garante consistência e enriquecimento semântico, mas também facilita a agregação de dados provenientes de múltiplas fontes. De maneira semelhante, o Cidoc CRM (Conceptual Reference Model) se distingue por sua abordagem centrada na representação semântica e nas relações complexas entre entidades do patrimônio cultural. Por carregar uma estrutura mais formal de classes, propriedades e relacionamentos bem definidos, permite o uso em raciocínio lógico e inferência sobre objetos culturais, eventos históricos, pessoas e lugares. Essa característica aproxima o Cidoc CRM do conceito de uma ontologia no domínio do patrimônio cultural.

Embora cada modelo tenha seu foco específico, todos compartilham objetivos fundamentais: assegurar a consistência das informações, facilitar a interoperabilidade e promover o intercâmbio eficaz de dados entre diferentes instituições e sistemas. Cabe lembrar, muitos softwares livres de gestão de acervos digitais, como o CollectiveAccess, ArchivesSpace e AtoM, já incorporam esses modelos de dados como *templates* de descrição, facilitando a adoção de práticas padronizadas por instituições que buscam preservar e disponibilizar suas coleções de forma estruturada e acessível.

Além da descrição de recursos, os repositórios também criam metadados relacionados à administração, ao acesso, à preservação e ao uso de coleções. Registros de aquisição, informações relacionadas a processos do sistema, informações técnicas de digitalização, controle de versões do objeto e acordos de licenciamento são exemplos desses outros tipos de metadados, em geral restritos à gestão interna do acervo.

Intimamente ligadas aos metadados estão as ontologias. Na essência, ontologias são estruturas conceituais que representam um domínio de conhecimento. Constituem a artéria central da web semântica e viabilizam a criação de modelos que vão de simples descrições de recursos até esquemas de classificação complexos.

Já mencionamos brevemente a Cidoc CRM, que é uma ontologia de domínio do patrimônio cultural, mas vamos esmiuçar um pouco mais esse construto.

Desde o início dos anos 1990, as ontologias têm sido objeto de estudos no interior de alguns campos da inteligência artificial, como engenharia do conhecimento, processamento de linguagem natural e representação do conhecimento (Studer et al., 1998). A partir da década seguinte, outras disciplinas também tomaram para si esse estudo, visando aplicações nas áreas de integração de sistemas, gestão do conhecimento, recuperação da informação e descoberta do conhecimento. Os primeiros trabalhos científicos sobre ontologias, em especial na apropriação de ontologias leves,[13] possuíam forte vínculo com as linguagens documentárias, chegando mesmo a haver pouca discriminação na terminologia do uso junto aos tesauros, pelas suas semelhanças de estrutura – ambas organizam conceitos e relações de um domínio – e de finalidade – para fins de representação e recuperação da informação (Moreira et al., 2003). No entanto, em termos de linguagem e nível de formalização existe uma diferença entre os dois instrumentos, já que tesauros cumprem o propósito de comunicação entre o usuário e as linguagens documentárias, e as ontologias o de descrição de objetos digitais para inferências computacionais.

No contexto de arquivos histórico-culturais, elas oferecem um mapa semântico que permite navegar pela complexidade das relações entre pessoas, eventos, lugares e documentos. No caso hipotético de um sistema que abriga conteúdos digitais referentes aos povos originários do Brasil, uma ontologia poderia ser utilizada para estruturar os dados relativos a práticas culturais, tradições orais, festas, e rituais desses povos. Dessa forma, a aplicação poderia relacionar documentos e objetos de maneira semântica, permitindo que pesquisadores e usuários finais explorem o acervo de forma mais profunda e contextualizada. Isso possibilitaria, por exemplo, a busca e análise de documentos segundo critérios como grupo étnico, localização geográfica, tipo de prática cultural ou período histórico, fornecendo uma visão interconectada e rica do patrimônio cultural dos povos indígenas do país.

Na prática, como tudo isso se articula? Em geral, uma ontologia é estruturada a partir de cinco elementos principais: conceitos, relações, propriedades, instâncias e axiomas. Os conceitos, ou classes, definem as categorias principais do domínio, como *povo indígena*, *tradição cultural*, *língua*, *ritual*, *artefato*. As relações estabelecem como esses conceitos se conectam, tal como a relação

13 As ontologias leves (lightweight ontologies) capturam relações básicas e conceitos essenciais dentro de um domínio específico, mas sem a complexidade e a formalidade das ontologias completas.

"realiza", que liga um *povo indígena* a um *ritual*. Já as propriedades de dados descrevem atributos específicos dos conceitos, como o "nome" de um povo ou "localização geográfica". As instâncias são exemplos concretos desses conceitos, como o povo *apinajé* ou o *ritual da corrida da tora*.[14] Por fim, axiomas são regras que definem restrições ou características adicionais, como "todo *ritual* deve ter pelo menos um *artefato* associado" ou "um *povo indígena* deve estar associado a pelo menos uma *língua*".

No entanto, ter ontologias bem definidas e metadados estruturados não é suficiente se a informação não puder ser efetivamente compartilhada. É aqui que entram os protocolos de comunicação. Estes são as "regras de etiqueta" do mundo digital, definindo como sistemas diferentes podem trocar informações de maneira padronizada, estabelecendo como os dados devem ser formatados, transmitidos, recebidos e interpretados. Implementar protocolos de comunicação permite que os sistemas façam solicitações e troquem informações de forma segura.

O OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting), por exemplo, é um protocolo de comunicação que utiliza um formato padronizado para expor metadados, geralmente no formato Dublin Core, o que facilita a interoperabilidade entre sistemas que podem ter diferentes estruturas internas de dados. Ele opera através de um modelo de provedor de dados e provedor de serviços: os provedores de dados disponibilizam metadados dos seus acervos, enquanto os provedores de serviços coletam e utilizam esses metadados para fornecer acesso aprimorado – por exemplo, por meio de motores de busca ou plataformas de agregação de dados – e serviços adicionais (Garcia; Sunye, 2003).

O "beabá" da interoperabilidade: enfrentando os jargões técnicos

No cenário atual de rápida evolução tecnológica, os centros de memória e instituições de guarda de acervos históricos enfrentam o desafio de não apenas preservar o passado, mas também de torná-lo acessível e relevante para o futuro. Nesse contexto, a interoperabilidade se apresenta como um conceito fundamental, embora frequentemente obscurecido por jargões técnicos. Em sua essência, a interoperabilidade refere-se à capacidade de diferentes sistemas e organizações trabalharem juntos de forma eficaz, compartilhando informações e conhecimentos (Marcondes, 2021, p. 19).

14 Há documentação sobre o patrimônio cultural do povo apinajé no acervo do Cpdoc. Para buscar no acervo, acesse: https://www18.fgv.br/cpdoc/acervo/arquivo.

Não é possível falar de interoperabilidade sem mencionar os princípios Fair (do inglês, *findability, accessibility, interoperability and reuse*), que visam tornar os dados mais encontráveis, acessíveis, interoperáveis e reutilizáveis.[15] Tornar itens de acervo mais encontráveis significa colocá-los na web, com metadados ricos, utilizando identificadores únicos e persistentes, como URLs fixas ou o identificador digital DOI, por exemplo. Para que esses itens sejam considerados acessíveis, eles devem estar indexados em repositórios digitais confiáveis, que garantam sua preservação a longo prazo. A acessibilidade também pode variar: alguns itens podem exigir cadastro ou pagamento, enquanto outros permitem o *download* gratuito e em alta resolução, incluindo seus metadados. A acessibilidade ideal, no entanto, vai além da simples disponibilidade; um item é ainda mais acessível quando é legível tanto por pessoas quanto por computadores, facilitando sua integração em sistemas automatizados e sua reutilização em diversos contextos, o que nos traz para a presente discussão sobre interoperabilidade e reuso.

A aderência a esses princípios garantirá que dados de diferentes fontes possam ser integrados e utilizados de maneira eficiente. Isso é particularmente importante em áreas como a ciência e a gestão de acervos digitais, nas quais a capacidade de compartilhar dados de forma confiável pode acelerar descobertas, enriquecer contextos e aprimorar a preservação do patrimônio cultural. Para as Glams, essa integração representa mais do que uma mera conveniência técnica; ela é a chave para desbloquear o verdadeiro potencial das coleções digitais. Imagine um pesquisador capaz de cruzar informações de manuscritos de um acervo com fotografias de outra instituição e registros audiovisuais de uma terceira, tudo isso sem sair de sua estação de trabalho. Essa não é uma visão futurista, mas uma realidade tangível que a interoperabilidade pode proporcionar.

No entanto, implementar a interoperabilidade em acervos digitais apresenta desafios únicos. A heterogeneidade dos materiais – que vão desde cartas manuscritas a gravações de história oral – exige abordagens flexíveis e robustas. O caminho começa com uma avaliação crítica de nossos sistemas existentes. Implica repensar como descrevemos e categorizamos os documentos e demais itens diante da necessidade de definir os padrões que permitirão a troca e o uso de dados entre diferentes sistemas. Isso inclui a adoção de formatos de dados abertos e esquemas de metadados padronizados, como Dublin Core, Mets, ou EAD, que garantem que os dados possam ser facilmente compartilhados. Em seguida, é crucial mapear e alinhar os metadados existentes nos acervos para

15 A Go-Fair Brasil é uma iniciativa que, alinhada à Go Fair global, promove os princípios Fair. A OpenGlam também é uma rede dedicada à promoção do acesso aberto em galerias, bibliotecas, arquivos e museus.

que sejam compatíveis com esses padrões, o que pode exigir sua conversão ou seu enriquecimento. Essa etapa pode ser complexa devido à diversidade de formatos e estruturas de dados presentes nos sistemas legados.

Desafios: um pequeno estudo de caso no Cpdoc

E quando há necessidade de realizar a migração de metadados a partir de um sistema legado? Essa tarefa pode ser bastante complexa e está longe de ser isenta de problemas. Primeiramente, pode haver inconsistências nos dados, como campos duplicados ou informações obsoletas, que precisam ser resolvidas antes da migração. Além disso, diferenças nos padrões de metadados entre o sistema antigo e o novo podem exigir mapeamentos detalhados e personalizações para garantir que todas as informações relevantes sejam preservadas e adequadamente interpretadas. Esses são desafios que nós também enfrentamos no Cpdoc, como discutiremos a seguir.

O sistema de busca no acervo do Cpdoc, nomeado Accessus,[16] foi lançado em 2001, utilizando tecnologias da Oracle, que já era uma das principais empresas de banco de dados no mundo, ao lado da Microsoft e IBM. Entre 2009 e 2010, alinhando-se às diretrizes internas de modernização tecnológica, o sistema foi migrado para o Microsoft SQL Server (MS SQL), com suporte aprimorado a consultas complexas, segurança avançada e ferramentas de integração e análise de dados. Para o usuário final, essa atualização resultou na implementação de um sistema de busca unificado, permitindo a consulta simultânea aos itens documentais do Programa de História Oral[17] (PHO), do Programa de Arquivos Pessoais[18] (PAP) e aos verbetes do Dicionário Histórico-Biográfico Brasileiro[19] (DHBB). Nessa oportunidade, foi também introduzida a ferramenta Colabore, que permitiu a interação da comunidade com a equipe do Cpdoc através do envio de informações complementares ou correções diretamente nos itens documentais.

Entre 2015 e 2016, o sistema passou por importantes atualizações, incluindo a remoção da obrigatoriedade de cadastro de usuário para acessar o acervo. Foram

16 Accessus. Disponível em: https://www18.fgv.br/CPDOC/acervo/arquivo-pessoal. Acesso em: 15 ago. 2024.

17 Programa de História Oral. Disponível em: https://cpdoc.fgv.br/acervo/historia-oral. Acesso em: 15 ago. 2024.

18 Programa de Arquivos Pessoais. Disponível em: https://cpdoc.fgv.br/acervo/arquivos-pessoais. Acesso em: 15 ago. 2024.

19 Dicionário Histórico-Biográfico Brasileiro. Disponível em: https://cpdoc.fgv.br/acervo/dicionarios/dhbb. Acesso em: 15 ago. 2024.

implementadas URLs amigáveis, que tornam os itens do acervo passíveis de serem indexados pelo Google e outras ferramentas de busca na internet. Além disso, foi possibilitado o acesso ao acervo através do aplicativo da FGV, instituição mantenedora, disponível para dispositivos Android e iOS. Todas essas novidades representaram um avanço significativo na abertura das coleções do Cpdoc ao público.

Do ponto de vista da catalogação arquivística, o Cpdoc foi pioneiro no desenvolvimento de uma metodologia para a descrição e organização de arquivos pessoais e entrevistas de história oral, inicialmente em áudio e posteriormente, também em vídeo. Desde os anos 1980, uma série de publicações da instituição detalhou os procedimentos técnicos adotados, tornando-os acessíveis ao público. A norma internacional Isad(G), aplicável a todos os tipos de materiais arquivísticos e utilizável em sistemas manuais e automatizados, foi lançada apenas em 1994. A versão em português, publicada pelo Arquivo Nacional, só foi disponibilizada em 1998.

Em outras palavras, o Cpdoc já estava desenvolvendo métodos para descrever e organizar seu acervo antes que normas e padrões internacionais fossem amplamente estabelecidos. Seus pesquisadores fizeram parte dos grupos que tiveram um papel fundamental na adequação desses padrões emergentes ao contexto nacional. Quando, em 2006, a norma brasileira de descrição arquivística Nobrade foi publicada, o Arquivo João Goulart, do Cpdoc, era um dos exemplos de aplicação da norma. Essa adoção por parte da instituição, no entanto, não se estende a todo o acervo e nem é totalmente precisa. Foram realizadas uma série de adaptações e escolhas alternativas que respondiam aos desafios e necessidades da instituição naquele momento.

Quadro 1 — Exemplo de equivalência entre metadados de identificação do item documental segundo as normas do Isad(G), Nobrade e Cpdoc

Isad(G)	Nobrade	Cpdoc manuscritos	Cpdoc audiovisual
Código(s) de referência	Código de referência	Classificação	Classificação
Título	Título	Fundo (nome do arquivo); Série (título); Subsérie (título)	Fundo (nome do arquivo), Título
Data(s)	Data(s)	Datas-limite do fundo; Datas-limite (de cada série); Período manuscrito; Ano de (manuscrito); Ano até (manuscrito); Precisão da data	Datas-limite do fundo; Datas-limite (de cada série); Período audiovisual; Ano de (audiovisual); Ano até (audiovisual); Precisão da data

Nível de descrição	Nível de descrição	Série, Subsérie	Série, Subsérie
Dimensão e suporte	Dimensão e suporte	Total de documentos do fundo; Quantidade de documentos; Quantidade (folhas, pastas ou páginas com informação)	Total de documentos do fundo; Quantidade documentos;

Fonte: elaborado pelas autoras, 2024.

No Quadro 1, apresentamos os elementos de metadados utilizados para identificar um item documental em quatro exemplos, conforme as normas Isad(G), Nobrade e as adaptações do Cpdoc, diferenciando entre itens textuais e audiovisuais. Vale destacar que as referidas normas gerais definem as informações de descrição (unidades semânticas) de itens e coleções, mas não especificam como os metadados devem ser representados em um sistema. Em outras palavras, elas não determinam os elementos de metadados em si. Além disso, para garantir a interoperabilidade, os rótulos dos campos em um banco de dados não precisam coincidir exatamente com os de um esquema padrão de metadados. É muito comum, aliás, que modelos de dados de uma mesma instituição combinem mais de um padrão de metadados.

Para a interoperabilidade acontecer, o que realmente importa é que haja uma correspondência clara entre os campos do banco de dados e os elementos do esquema padrão, mesmo que os rótulos sejam diferentes. Essa correspondência pode ser estabelecida por meio de mapeamentos, como demonstrado no Quadro 2.

Quadro 2 — Exemplo de correspondência entre elementos da base Accessus (Cpdoc) e dos esquemas Dublin Core e EAD

Elemento da norma	Metadado (visível pelos usuários internos e externos)	Campo da base de dados (exemplo de audiovisual)	Dublin Core	Encoded Archival Description EAD
Código de referência	Fundo, Série, Subsérie, Complemento (interno); Classificação (externo)	SG_FUN, SG_ SIGLA_SER, CD_ CLASSIFICACAO_AVI	dc:identifier	<unitid>
Título	Título	DS_TITULO_AVI	dc:title	<unittitle>

Fonte: elaborado pelas autoras, 2024.

Um exercício de mapeamento de metadados entre um conjunto pequeno de elementos do acervo do Cpdoc[20] e o Dublin Core mostrou, na prática, o tamanho do desafio colocado por esse tipo de mapeamento. Notamos que dados originalmente separados no esquema do Accessus por vezes correspondiam a um único elemento no Dublin Core, evidenciando a necessidade de consolidação, como já exemplificado no Quadro 2. Esse também foi o caso dos metadados dc:date e dc:format, que correspondiam a mais de um metadado no esquema do Cpdoc. Além disso, certos metadados presentes no Dublin Core não encontravam equivalentes diretos no esquema original do Cpdoc, criando lacunas de correspondência, como no caso de dc:rights para os direitos de acesso a nível do item documental. Ademais, a URL de cada recurso em nosso acervo não era registrada no banco de dados, o que impedia, a princípio, sua inclusão no metadado dc:identifier.

Outro problema identificado foi a inconsistência na nomenclatura dos metadados, com divergências nos nomes dos campos utilizados nas interfaces interna e externa do sistema. Assim, o mesmo elemento de descrição recebia um rótulo na tela de cadastro de itens e outro diferente na tela de navegação do usuário. Para esclarecer essas discrepâncias, realizaram-se consultas SQL na base de dados do Cpdoc, permitindo verificar a nomenclatura correta dos metadados.

Esses exercícios demonstraram a importância de um planejamento cuidadoso na elaboração do modelo de dados de uma coleção. Assegurar a interoperabilidade das coleções significa que os dados e as informações devem ser compreendidos e utilizados de forma consistente em diferentes sistemas. Isso envolve a dimensão semântica (modelos conceituais e vocabulários), a dimensão sintática (estrutura dos metadados) e a dimensão tecnológica (linguagens de marcação como XML, RDF ou JSON). Além disso, a qualidade da descrição das informações é crucial para o desenvolvimento de aplicações de *data science*, pois melhora a extração (obtenção de dados de diferentes fontes), a geração (criação de novos *insights*) e o reuso (aplicação dos dados em novos contextos).

20 A amostra para o teste de integração dos acervos na infraestrutura Rossio contou com cinco entrevistas, dez manuscritos e dez fotografias com temas relacionados a Portugal.

Novos tempos e novas abordagens: a inteligência artificial em cena

Nos últimos anos, o Cpdoc tem se dedicado à exploração de metodologias digitais inovadoras para modernizar e ampliar o acesso ao seu acervo histórico. Tecnologias avançadas de processamento de dados, com uso da inteligência artificial e de aprendizado profundo, têm transformado tarefas tradicionalmente laboriosas, como a transcrição de textos manuscritos (OCR), a legendagem, e a classificação e sumarização de documentos textuais e audiovisuais, agilizando o processo de catalogação e aprimorando os pontos de acesso às informações. No entanto, o alto custo dessas tecnologias e o conhecimento especializado exigido para sua implementação ainda limitam seu uso, que permanece em fase exploratória e abrangendo conjuntos documentais reduzidos. Apesar disso, já é possível antever o potencial que essas inovações têm para redefinir as fronteiras do trabalho de curadoria digital.

Até recentemente, o reconhecimento dos textos manuscritos das nossas coleções digitalizadas se traduzia por camadas de OCR sem quaisquer chances de aproveitamento, como mostra o exemplo da Figura 3. A imagem da página contendo anotações escritas à mão integra o arquivo privado do antropólogo Gilberto Velho, depositado no Cpdoc, e a versão "reconhecida" desta página — fornecida pela empresa de digitalização —, não passa de uma sequência sem sentido de caracteres.

Com as ferramentas de OCR aprimoradas pela IA e pelos Modelos de Linguagem em Larga Escala (LLMs), os resultados melhoraram significativamente, como mostrado na Figura 4, na qual o reconhecimento das palavras foi praticamente perfeito. Além disso, esses modelos, treinados em vastos *corpora* textuais, conseguem aprender padrões linguísticos complexos, gerando textos coerentes e contextualmente relevantes em tarefas como descrição e sumarização automáticas. No exemplo citado, além do OCR, o modelo foi instruído a gerar uma breve descrição da imagem e a identificar as entidades mencionadas no texto reconhecido. O resultado foi exatamente este: "Descrição: A página de um caderno manuscrito contendo anotações detalhadas sobre a descrição de um espaço habitacional, possivelmente parte de uma pesquisa ou estudo antropológico. O texto descreve o layout e as características de uma moradia, incluindo detalhes sobre cômodos, ocupação e condições de aluguel". As entidades reconhecidas foram todas do tipo *lugar*: cozinha, banheiro, sala, quarto conjugado.

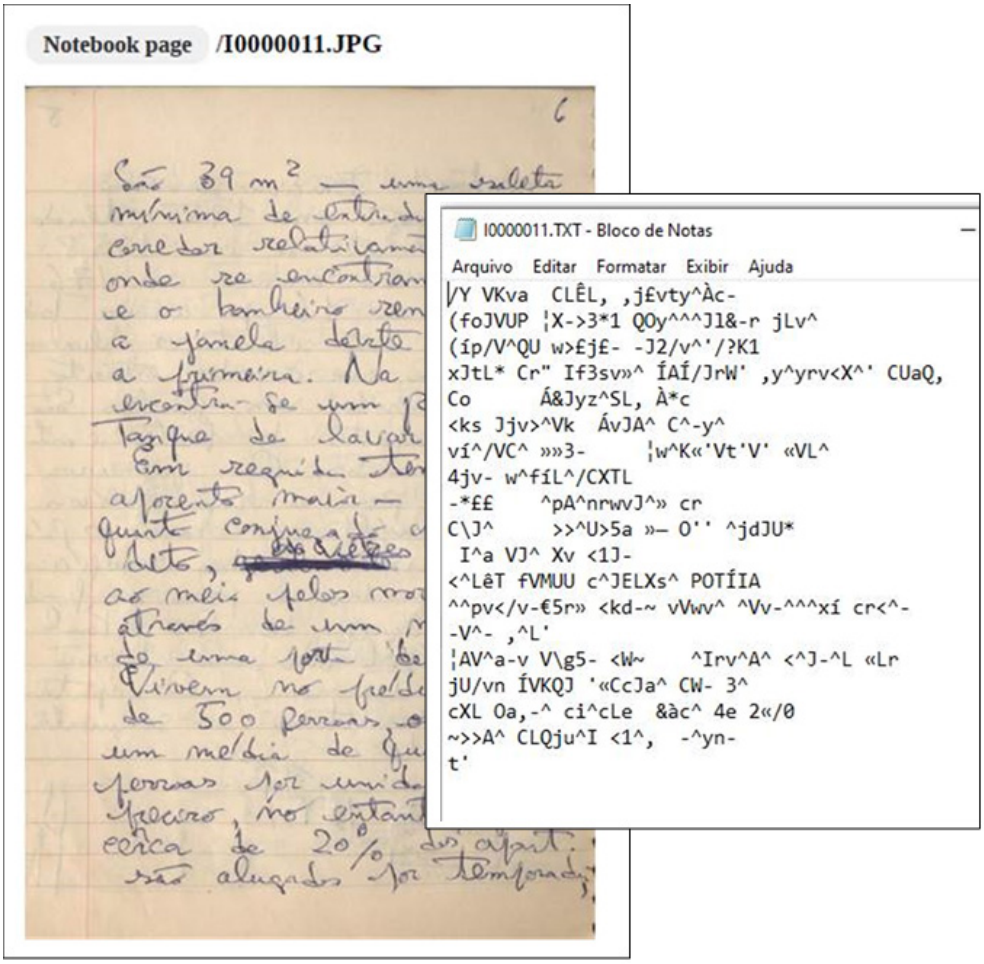

Figura 3 — Página manuscrita de um documento de Gilberto Velho e o texto obtido no Reconhecimento Óptico de Caracteres (OCR) pela empresa de digitalização. Fonte: elaborada pelas autoras, 2024

São 39 m² - uma saleta
mínima de entrada, um
corredor relativamente longo
onde se encontram a cozinha
e o banheiro sendo que
a janela deste dá para
a primeira. Na cozinha
encontra-se um pequeno
tanque de lavar roupa.
Em seguida temos o
aposento maior - o sala-
quarto conjugado efetivamente
dito, que é dividido
ao meio pelos moradores
através de um móvel ou
de uma lona. Se juntarmos
vivem no prédio por volta
de 500 pessoas, o que dá
um média de quase 3
pessoas por unidade. É
preciso, no entanto, dizer que
cerca de 20% dos apart.
são alugados por temporada,

Figura 4 — Texto obtido utilizando o modelo de linguagem Claude 3.5 Sonnet. Fonte: modelo construído pelo bolsista e cientista de dados Álvaro Justen

A transcrição automática de entrevistas também alcançou avanços notáveis. Modelos de reconhecimento de fala baseados em aprendizado de máquina, como o openAI Whisper, agora oferecem resultados excepcionais, permitindo uma economia significativa de tempo e recursos nessa tarefa — ainda que não exclua o trabalho humano, necessário para a validação rigorosa dos resultados e para as eventuais correções que se façam necessárias. Utilizando o modelo Whisper, processamos uma seleção de entrevistas do acervo do Cpdoc e geramos transcrições a partir dos áudios. Para cada entrevista, já contávamos com transcrições feitas manualmente, portanto já corrigidas e revisadas pela equipe. O objetivo então, no nosso caso, foi criar um algoritmo capaz de alinhar os *timestamps* das transcrições automáticas com o texto das transcrições manuais, de qualidade superior. O resultado pode ser visualizado na interface desenvolvida,[21] que exibe os vídeos das entrevistas com legendas sincronizadas e inclui um campo de pesquisa de termos e palavras-chave que direciona diretamente aos trechos específicos onde esses termos são mencionados (Figura 5).

21 Disponível em: https://cpdoc.fgv.br/cientistassociais/explore.

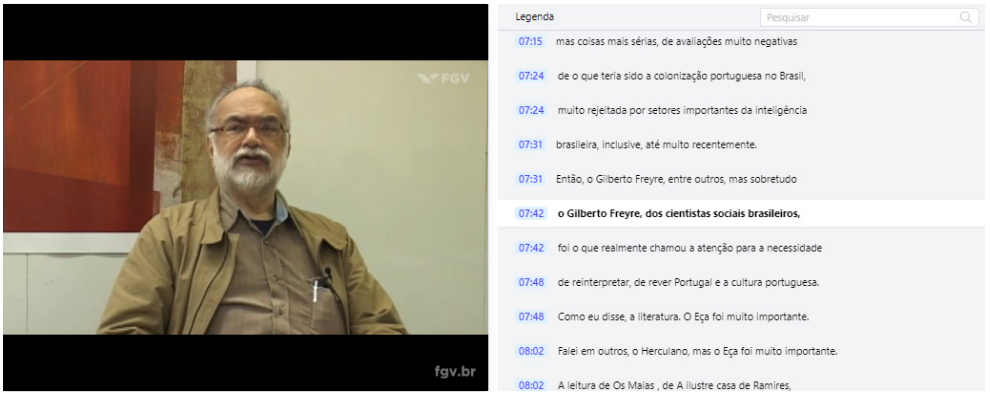

Figura 5 — Interface exibindo entrevista com áudio sincronizado à transcrição já revisada e corrigida. Fonte: elaborada pelas autoras, 2024

Outros avanços também se destacam na área de curadoria digital, especialmente no processamento de imagens. A aplicação de redes neurais e técnicas de reconhecimento de padrões tem revolucionado a restauração de fotografias históricas, permitindo a remoção de ruídos indesejados, o aprimoramento da nitidez e a correção de variações de iluminação, resultando em imagens de alta qualidade que preservam suas características históricas essenciais (Higuchi; Rocha, 2024). Além disso, essas tecnologias facilitam a identificação de elementos específicos nas imagens, como rostos, objetos e cenários, permitindo uma catalogação mais precisa e uma melhor integração com bancos de dados históricos. Isso não apenas enriquece a preservação dos acervos, mas também expande as possibilidades de pesquisa e interpretação dos materiais digitalizados.

Também é importante mencionar os bancos de dados semânticos. Embora tradicionalmente construídos com base em ontologias, eles têm se beneficiado de novas alternativas proporcionadas pelos avanços nas técnicas de *machine learning*. Atualmente, os *vector databases* oferecem uma abordagem inovadora, utilizando representações vetoriais para objetos de texto e imagens (Masoudi, 2024). Em vez de depender do alinhamento rígido de vocabulário característico das ontologias, a busca semântica nesses sistemas baseia-se na similaridade entre vetores. Isso permite uma recuperação de informações mais fluida e adaptável, pois a similaridade vetorial pode capturar nuances de significado que o alinhamento de vocabulário pode não abranger tão bem. Embora essa abordagem simplifique a recuperação de informações em muitos casos, ela ainda não atinge a integração ao nível do *linked open data*, que requer a adoção de descrições mais formais e ontologias complexas. A solução baseada em vetores oferece uma flexibilidade valiosa, mas a integração completa com o ecossistema de dados aberto e interconectado ainda demanda um maior esforço e o uso de modelos mais estruturados.

Em suma, essas novas abordagens prometem elevar a acessibilidade dos patrimônios documentais a patamares sem precedentes. A integração com ferramentas de inteligência artificial e análise de dados não só contribui com a democratização do acesso, mas também facilita e acelera a pesquisa acadêmica, permitindo análises em larga escala e a identificação de padrões e conexões antes difíceis de detectar.

Em Marques (2024), encontramos um exemplo de trabalho que explora o uso de técnicas computacionais, especificamente a modelagem de tópicos e o reconhecimento automático de entidades nomeadas, para aprimorar a pesquisa de arquivos históricos. A autora aplicou essas metodologias ao Arquivo Juarez Távora, depositado no Cpdoc, utilizando mineração de texto com Python e análise qualitativa. A modelagem de tópicos permitiu identificar e avaliar a relevância de assuntos no acervo, selecionando automaticamente 2.601 arquivos digitais (4,8% do total) relacionados ao tema "Brasil rural". O reconhecimento de entidades nomeadas também foi importante para compreender o papel de novos atores, como o movimento tenentista e a nova burocracia. A autora destaca que a combinação de leitura distante (análise computacional) e próxima (análise qualitativa detalhada) é desafiadora, mas valiosa, permitindo uma visão abrangente dos temas predominantes e selecionando material relevante para uma análise mais profunda, que captura nuances de estilo, personalidade e emoções nos documentos.

Assim, as oportunidades proporcionadas pelas novas metodologias digitais aplicadas a arquivos histórico-documentais não apenas promovem a interdisciplinaridade, mas também catalisam o surgimento de campos de estudo inovadores. Essa convergência entre tecnologia e patrimônio documental enriquece significativamente o panorama acadêmico e cultural, abrindo novas perspectivas para a pesquisa e interpretação histórica.

Considerações finais

A visão de uma internet mais democrática, descentralizada e transformadora é um projeto ainda em construção. A «datificação" de acervos históricos e culturais contribui para esse projeto ao redefinir a forma como interagimos com nosso patrimônio. Esse processo, que vai além da simples digitalização, transforma fontes primárias em conjuntos de dados estruturados, abrindo um universo de possibilidades. Com essa conversão, é possível aplicar técnicas avançadas de análise de dados e aprendizado de máquina aos arquivos, revelando padrões e conexões antes invisíveis. A interoperabilidade entre coleções cria um rico ecossistema digital de conhecimento, enquanto o acesso facilitado convida tanto pesquisadores quanto o público em geral a explorar o passado e o presente de formas inovadoras.

Essas técnicas nos permitem alcançar vários objetivos úteis, como a classificação automática e a análise estatística dos temas presentes no acervo, ou ainda o enriquecimento de metadados, por meio do reconhecimento e da extração automática de entidades, o que aprimora a organização e a pesquisa dos itens digitais. Além disso, há ferramentas que possibilitam a realização de análises de redes e relações, com a identificação de conexões e padrões entre diferentes elementos do acervo, a abertura de perspectivas inéditas sobre o material e a facilitação de descobertas que, de outra forma, seriam difíceis de perceber.

A intrincada transformação digital que discutimos neste artigo está redefinindo profundamente a natureza do trabalho em instituições de acervo cultural e histórico. Trata-se de um processo sociotécnico que não apenas introduz novas ferramentas e metodologias, mas também reconfigura as competências necessárias e as dinâmicas de interação entre equipes, com o público e com o patrimônio cultural. Existe uma crescente demanda, talvez ainda não vocalizada em alto e bom som, por profissionais com perfis híbridos, que combinem conhecimentos tradicionais das ciências da informação e das ciências humanas com habilidades técnicas em áreas como ciência de dados, programação e design de experiência do usuário.

É urgente que pensemos em trilhas formativas que atendam a essas necessidades e que passemos a trabalhar em equipes interdisciplinares e mesmo em redes multi-institucionais, com o respeito, a paciência e a curiosidade que as relações com a alteridade impõem. Não podemos perder de vista que somente uma abordagem interdisciplinar oferecerá o ferramental teórico e prático necessário para viabilizar inovações na curadoria, preservação e disseminação de acervos digitais. Novas formas de engajamento com as coleções, desde visualizações interativas e análises computacionais em larga escala até experiências de realidade virtual e aumentada, precisam deixar de ser o futuro e começar a se materializar no presente.

Essas mudanças, enfim, não apenas ampliam o alcance e o impacto das coleções, mas também desafiam as instituições a repensarem seus papéis como mediadoras culturais na era digital e frente a desafios relacionados a desigualdades sociais e de acesso a recursos financeiros. Por isso, trabalhar em colaboração nos parece fundamental. Ao compartilhar nossas descobertas e reflexões, esperamos contribuir para o diálogo mais amplo sobre a gestão e disseminação eficazes de coleções digitais e inspirar novas abordagens neste campo em constante evolução.

Referências

BACA, Murtha. Introduction. In: BACA, Murtha (org.). *Introduction to metadata*. 3. ed. Los Angeles: Getty Publications, 2016. Disponível em: https://www.getty.edu/publications/intro-metadata/. Acesso em: 20 nov. 2023.

BOITA, Tony; BAPTISTA, Jean. Presença e resistência: memórias LGBTQIA+ nos museus brasileiros - Apresentação do Dossiê "Memória, Museologia LGBTQIA+ e museus nacionais". *Anais do Museu Histórico Nacional*, Rio de Janeiro, v. 58, p. 1-2, 2024. Disponível em: https://anaismhn.museus.gov.br/index.php/amhn/article/view/342. Acesso em: 20 nov. 2023.

BOUGHIDA, Karim. CDWA lite for Cataloguing Cultural Objects (CCO): A new XML schema for the cultural heritage community. In: INTERNATIONAL CONFERENCE OF THE ASSOCIATION FOR HISTORY AND COMPUTING, 16., 2005, Amsterdam. *Proceedings*: Humanities, Computers and Cultural Heritage. p. 14-17.

BROWN, Karen; CUMMINS, Alissandra; GONZÁLEZ RUEDA, Ana S. (org.). *Communities and museums in the 21st century*: shared histories and climate action. 1. ed. Londres: Routledge, 2023. Disponível em: https://doi.org/10.4324/9781003288138. Acesso em: 20 nov. 2023.

DAQUINO, M., Tomasi, F. Historical Context Ontology (HiCO): a conceptual model for describing context information of cultural heritage objects. In: GAROUFALLOU, E.; HARTLEY, R.; GAITANOU, P. (ed.). *Metadata and semantics research*. MTSR 2015. Communications in Computer and Information Science, v. 544. Springer, 2015. Disponível em: https://doi.org/10.1007/978-3-319-24129-6_37. Acesso em: 20 nov. 2023.

FOUCAULT, Michel. *O corpo utópico, as heterotopias*. São Paulo: Edições, 2013.

GARCIA, Patrícia de Andrade Bueno; SUNYE, Marcos Sfair. *O protocolo OAI-PMH para interoperabilidade em bibliotecas digitais*. In: CONGRESSO DE TECNOLOGIAS PARA GESTÃO DE DADOS E METADADOS DO CONE SUL, 4., 2003. Disponível em: https://conged.deinfo.uepg.br/. Acesso em: 20 nov. 2023.

GILLILAND, Anne J. Setting the stage. In: BACA, Murtha (ed.). *Introduction to metadata*. 3. ed. Los Angeles: Getty Publications, 2016. Disponível em: https://www.getty.edu/publications/intro-metadata/. Acesso em: 20 nov. 2023.

HIGUCHI, Suemi; SOUZA, Renato Rocha. Ciência de dados e arquivos. In: CASTRO, Celso; SPOHR, Martina; BLANK, Thais (org.). *Arquivos pessoais*: experiências no Cpdoc. v. 1. 1. ed. Rio de Janeiro: FGV Editora, 2024. p. 180-208.

JONES, Elisabeth. The public library movement, the digital library movement, and the large-scale digitization initiative: assumptions, intentions, and the role of the public. *Information & Culture*, v. 52, n. 2, p. 229-263, 2017. Disponível em: http://www.jstor.org/stable/44667555. Acesso em: 20 nov. 2023.

KNUDSEN, Britta Timm; OLDFIELD, John; BUETTNER, Elizabeth; ZABUNYAN, Elvan. (org.). *Decolonizing colonial heritage*: new agendas, actors and practices in and beyond Europe. 1. ed. Londres: Routledge, 2021. Disponível em: https://doi.org/10.4324/9781003100102. Acesso em: 20 nov. 2023.

LIDDINGTON, Jill. What is public history? Publics and their pasts, meanings and practices. *Oral History*, v. 30, n. 1, p. 83-93, 2002. Disponível em: http://www.jstor.org/stable/40179644. Acesso em: 20 nov. 2023.

MARCONDES, Carlos Henrique. *Dados abertos interligados*: publicação, recuperação e integração de acervos de arquivos, bibliotecas e museus na web. Botucatu: Oficina Universitária, 2021. Disponível em: https://doi.org/10.36311/2021.978-65-5954-040-2. Acesso em: 20 nov. 2023.

MARQUES, Juliana. Arquivos pessoais e elites políticas: uma abordagem digital para o estudo do Brasil rural na Era Vargas. In: CASTRO, Celso; SPOHR, Martina; BLANK, Thais (org.). *Arquivos pessoais*: experiências no Cpdoc. v. 1. 1. ed. Rio de Janeiro: FGV Editora, 2024. p. 142-179.

MARTINS, Dalton Lopes. *Acervos digitais nos museus*: manual para realização de projetos. Brasília: Instituto Brasileiro de Museus, 2020. 140p.

MASOUDI, Arian. *AggroDoc-semantic search using vector databases and large language models*. Thesis (Master programme in Computer Science and Engineering) – Luleå University of Technology, Luleå, Suécia, 2024.

MONTEIRO, Juliana. Compartilhamento de acervos na internet: reflexões a partir da prática. *Museologia & Interdisciplinaridade*, [s. l.], v. 10, n. especial, p. 61-72, 2021. Disponível em: https://periodicos.unb.br/index.php/museologia/article/view/38213. Acesso em: 10 ago. 2024.

MOREIRA, Alexandra; ALVARENGA, Lídia; OLIVEIRA, Alcione Paiva. O nível do conhecimento e os instrumentos de representação: tesauros e ontologias. *DataGramaZero*, João Pessoa, v. 5, n. 6, 2004.

OPEN KNOWLEDGE FOUNDATION. *Open Data Handbook*. [s. l.]: [s. n.], 2012. Disponível em: http://opendatahandbook.org. Acesso em: 5 ago. 2024.

SÃO PAULO (Estado). Secretaria de Orçamento e Gestão. Arquivo Público do Estado de São Paulo Política de gestão e preservação de documentos digitais. São Paulo: Arquivo Público do Estado de São Paulo, 2022. Disponível em: http://www.arquivoestado.sp.gov.br/uploads/publicacoes/livros/politica_de_gestao_e_preservacao_de_documentos_digitais.pdf. Acesso em: 5 ago. 2024.

SOUZA, R. R.; COELHO, F.; HIGUCHI, S.; SILVA, D. L. Pesquisas em organização de informação na FGV: o portal semântico do Cpdoc. In: Isko Brasil, 1., 2011. *Anais...* Brasília: Isko Brasil, 2011. p. 227-232. Disponível em: http://hdl.handle.net/20.500.11959/brapci/135041. Acesso em: 5 ago. 2024.

STUDER, Rudi; BENJAMINS, Richard; FENSEL, Dieter. Knowledge engineering: principles and methods. *Data and Knowledge Engineering*, 25, 1998.

TIC Domicílios – Pesquisa sobre o uso das tecnologias de informação e comunicação nos domicílios brasileiros: TIC Domicílios 2022. Núcleo de Informação e Coordenação do Ponto BR (ed.). São Paulo: Comitê Gestor da Internet no Brasil, 2023.

TIC Cultura – Pesquisa sobre o uso das tecnologias de informação e comunicação nos equipamentos culturais brasileiros: TIC Cultura 2022. Núcleo de Informação e Coordenação do Ponto BR (ed.). São Paulo: Comitê Gestor da Internet no Brasil, 2023.

Recebido em 24/8/2024
Aprovado em 13/3/2025