



Foto: Freepik.com. Reprodução

A digitalização é fundamental para preservar e manter dados históricos e inéditos.

Preservação e acesso a dados históricos e inéditos

por Bianca Bosso

Na manhã de 4 de agosto de 1944, pouco antes do fim da Segunda Guerra Mundial, um telefonema anônimo para uma unidade nazista de caça aos judeus resultou na prisão de oito pessoas escondidas em um edifício comercial na Holanda. Entre as vítimas, estava Anne Frank, uma garota de quinze anos que teve seus relatos sobre a guerra publicados pelo pai após falecer em um campo de concentração. Mais de 80 anos e duas investigações oficiais depois, o mistério sobre a identidade do delator anônimo ainda persiste. No entanto, avanços modernos na preservação e no acesso a dados históricos, como a digitalização de documentos

e o uso de inteligência artificial (IA), permitiram revisitar o caso recentemente, revelando detalhes inéditos e oferecendo novas perspectivas para uma das questões mais intrigantes da história.

Em meados de 2016, o cineasta belga Thijs Bayens transformou seu interesse pelo caso Anne Frank em um ambicioso projeto de investigação. Ele reuniu uma equipe com mais de vinte detetives, historiadores, pesquisadores e outros especialistas em torno da busca pelo delator, trabalho posteriormente denominado "Caso Arquivado". Ao longo de cinco anos, os profissionais analisaram milhares de documentos históricos, incluindo

registros policiais, listas de deportação, cartas, arquivos e outros vestígios espalhados por vários continentes. O estudo detalhado de tal quantidade de materiais poderia ter sido inviável em décadas passadas, mas com o apoio de ferramentas tecnológicas, se tornou possível. "Além de documentos e livros escaneados, a parte de reconhecimento por voz [...] converteu gravações de vídeo e áudio em texto, os tornou passíveis de buscas e os traduziu para o inglês" relata Rosemary Sullivan no livro *"Quem Traiu Anne Frank? A investigação que revela o segredo jamais contado"*, onde detalha a pesquisa.

Com as informações digitalizadas, ordenadas e disponíveis para acesso remoto, a equipe começou a notar conexões inéditas entre pessoas, lugares, datas e endereços. “Uma vez digitalizadas, você pode transformar as informações em dados; então, é possível cruzar informações e, eventualmente, submeter às tecnologias de reconhecimento de texto e de IA”, destaca Thiago Nicodemo, pesquisador e diretor do Arquivo Público do Estado de São Paulo. O potencial da nova estratégia foi rapidamente notado. “Se, por exemplo, um endereço interessante surgia em algum documento que eu estava examinando, eu podia cruzá-lo muito rapidamente com o banco de dados. A análise do endereço pelo programa de IA me fornecia todos os documentos e outras fontes relevantes onde o endereço era mencionado”, explica Pieter van Twisk, sócio de Thijs Bayens nas investigações, para Rosemary Sullivan. (Figura 1)

Digitalização de documentos para preservação e acesso

As experiências descritas por Rosemary Sullivan sobre o “Caso Arquivado” trazem à tona alguns dos aspectos da forma moderna de lidar com registros históricos. Documentos que, antes, compunham pilhas de papéis em caixas e prateleiras, podem, agora, ocupar apenas alguns *gigabytes* em um espaço virtual. Thiago Nicodemo ressalta que essa

transformação traz vantagens significativas, especialmente na superação de barreiras temporais, espaciais e econômicas para o acesso às informações. “Os dados ficam acessíveis para muito mais gente e pessoas que não necessariamente são especialistas podem acessar documentos históricos que só especialistas conseguiriam ver nos acervos”, afirma.

“A pesquisa que fazemos hoje e que faremos no futuro tem um perfil muito diferente do que foi feito no passado”

O pesquisador destaca ainda que o armazenamento de dados históricos em formato digital também implica em mudanças no perfil das pesquisas conduzidas. “Você pode responder perguntas que são inéditas, sob pontos de vista que não conseguia enxergar antes, porque, agora, consegue olhar uma perspectiva em larguíssima escala”, observa. O aspecto descrito por Thiago Nicodemo foi um ponto crucial para a trajetória do “Caso Arquivado”, uma vez que possibilitou observar as informações sob olhares diferentes dos empregados nas investigações anteriores, conduzidas em 1947 e 1963. “A pesquisa que fazemos hoje e que faremos no futuro tem um perfil muito diferente do que foi feito no passado”, prevê.

Junto aos benefícios, no entanto, as novas formas de armazenamento e organização

trazem consigo novos desafios. Rodrigo Esteves de Lima-Lopes, professor do Instituto de Estudos da Linguagem da Unicamp e pesquisador na área de linguagem e tecnologias, aponta uma dessas dificuldades. “Dados que não estejam consolidados e com seus metadados claramente organizados podem não ser encontrados ou mesmo utilizáveis por aqueles que deles precisam”, pontua. Thiago Nicodemo concorda e acrescenta que a infraestrutura de armazenamento também deve ser considerada. “Existe um risco altíssimo da perda desses registros digitais a longo prazo se não houver manutenção dessas estruturas. Além disso, às vezes, você cria documentos em linguagens e formatos que não consegue mais abrir depois de cinco ou 10 anos”, explica.

É preciso entender o objetivo antes de digitalizar

Considerando os prós e contras do processo, Thiago Nicodemo explica que a digitalização de dados pode ter dois objetivos principais. Em projetos de cruzamento e reunião de dados, como o liderado por Thijs Bayens, onde as informações estão dispersas por diversos países, em línguas e formatos variados, a digitalização facilita a disponibilização das informações. “Por exemplo, você vai em vários arquivos do mundo e digitaliza tudo que tem a ver com escravidão. Daí, o usuário vai ter acesso a uma experiência muito diferente do que se ele estivesse



Foto: APESP. Reprodução

Figura 1. Arquivo Público do Estado de São Paulo

pesquisando na estrutura do arquivo normal, porque pode se tornar especialista em um nível mais amplo, não apenas na região com acesso, por exemplo”, comenta, destacando que o uso consciente das novas tecnologias pode beneficiar uma grande diversidade de temas.

Em outros casos, o armazenamento virtual pode ser uma alternativa para preservar acervos muito antigos ou em degradação, evitando que as informações se percam com o tempo. “Nessas situações, a digitalização é uma forma fundamental de fazer com que aqueles dados se eternizem”, explica o pesquisador. “Mas, às vezes, vale mais a pena investir para cuidar dos papéis que estão ruindo do que investir uma grande quantia de dinheiro em um projeto de

digitalização e se perder pelo caminho”, complementa.

Apenas uma fração do que existiu é armazenada

No fim das investigações do “Caso Arquivado”, a equipe detinha mais de 66 *giga-bytes* de informações na forma de 7.500 documentos. O volume de informações, equivalente a 33 filmes modernos de duas horas em qualidade padrão, é, no entanto, apenas uma parte de todos os relatórios, diários, recibos, imagens e outros relatos gerados antes, durante e após a prisão do grupo de Anne Frank após a ligação anônima. Ainda assim, permitiu extrair e correlacionar informações riquíssimas.

Vincent Pankoke, ex-agente do Departamento

Federal de Investigação dos Estados Unidos (FBI) e líder da equipe “Caso Arquivado”, encontrou cerca de 956 formulários que identificavam a entrega de recompensas para policiais do Serviço de Segurança Alemão (SD) que encontraram e apreenderam judeus. Os registros, contudo, eram datados do início de 1942 a meados de 1943 e deixavam de fora possíveis pagamentos realizados além desse período. A perda desses dados pode ter acontecido por diversas razões. Em seu livro, Rosemary Sullivan explica que um bombardeio britânico em 1944 destruiu um prédio que guardava recibos administrativos. Além disso, quando a derrota ficou iminente para os nazistas, a tentativa de eliminar as provas sobre o Holocausto resultou na queima de milhares de documentos, exterminando para sempre

detalhes que completariam esse quebra-cabeça.

A incompletude de registros (digitais ou não) não é algo exclusivo da Segunda Guerra Mundial. Quaisquer arquivos que tenham sido preservados até os dias atuais representam apenas uma pequena porcentagem do que já existiu e simbolizam a existência de um mecanismo que determina, de forma intencional ou não, a história que será contada no futuro. “Por exemplo, logo depois que terminou a escravidão, ainda na década de 1890, Rui Barbosa ordenou a incineração de todos os documentos que tivessem a ver com a escravidão. É quase a regra”, observa Thiago Nicodemo. “Se um documento sobrevive até hoje, alguém lutou para guardar aquilo. Em qualquer mídia, tanto no papel quanto no digital”, complementa.

“Se um documento sobrevive até hoje, alguém lutou para guardar aquilo.”

Assim como em um projeto de coleta manual de informações, processos digitais também podem resultar em um enviesamento de dados e análises devem levar esse fator em conta. Rodrigo Lima-Lopes menciona, como exemplo, o caso das mídias sociais, que recebem, armazenam e circulam milhões de informações a cada instante. “Fatos contemporâneos nos mostram que as empresas privadas que gerenciam as mídias sociais tomam posturas que podem excluir a voz de

determinados grupos sociais e políticos, criando uma representação que não espelha, efetivamente, a sociedade”.

Como escolher o que será preservado?

Embora muito se fale, em Ciência Aberta, na necessidade de armazenar de forma aberta produtos da pesquisa, para compartilhamento, não basta pensar em armazenar. É preciso também considerar a preservação (e, inclusive, o decaimento digital e obsolescência dos dispositivos de armazenamento).

A preservação de registros pode ser influenciada por decisões políticas, como a escolha por destruir certos itens, e por fatores acidentais, como falhas no processo de apagamento ou de backup. Apesar das tentativas, muitos documentos entre 1939 e 1945 persistiram, incluindo as milhares de páginas investigadas pelo “Caso Arquivado”. Eventos como o bombardeio britânico de 1944 e catástrofes naturais também impactam o que é mantido. “A última bola da vez foi o incêndio em Los Angeles, mas também tivemos a pandemia, as enchentes no Rio Grande do Sul...”, exemplifica Thiago Nicodemo.

A disponibilidade de recursos para a preservação também é um fator decisivo. “Normalmente, [arquivos digitais] são guardados em servidores de instituições e essas instituições vivem instabilidades políticas no Brasil, no plano da cultura, por exemplo. Há muita inconsistência nos

investimentos”, diz Thiago Nicodemo. Uma alternativa atual tem sido enviar os dados para sistemas de nuvem, serviços online que permitem armazenar arquivos na internet, sem a necessidade de dispositivos físicos. No entanto, o pesquisador defende que essa pode não ser a melhor alternativa a longo prazo. “Se você fez um contrato de nuvem ou uma base de dados em um aplicativo e deixa de pagar depois de um tempo, você não consegue mais acessar o que produziu”, explica.

Os próprios projetos de manutenção de acervos e digitalização desempenham papel ativo na decisão do que será mantido. O diretor do Arquivo Público do Estado de São Paulo (APESP) relata que, devido aos custos de manutenção, acervos costumam digitalizar documentos com altas taxas de acesso em detrimento daqueles que são pouco procurados. Gilberto Lacerda dos Santos, que coordena o Museu Virtual de Ciência e Tecnologia da Universidade de Brasília, acrescenta que tais instituições ainda precisam considerar duas dimensões éticas: “uma é ligada à natureza, à qualidade e à veracidade do conteúdo e a outra, aos direitos autorais”, diz. (Figura 2)

Projetos particulares, como o “Caso Arquivado”, também podem atuar na digitalização de documentos sobre um único tema ou uma área de estudos de interesse em larga escala, mas isso pode trazer riscos para os acervos físicos responsáveis pelas versões originais dos registros. “O mais atraente é quando alguma instituição internacional oferece a



Foto: Marcelo Camargo/ Agência Brasil. Reprodução

Figura 2. Museus reúnem arte, ciência e tecnologia em acervos em ambiente virtual

digitalização desses documentos para fazer grandes projetos transversais. Isso até pode ser uma solução, mas, enquanto você colabora para aumentar a relevância dos dados, as instituições de origem que estão guardando os documentos se tornam menos relevantes”, explica Thiago Nicodemo.

No Brasil: preservação e acesso a dados sobre a ditadura

Assim como os documentos digitalizados pela equipe de Thijs Bayens contribuem para recontar a história da Europa, os esforços para preservar e ampliar o acesso a registros da ditadura militar brasileira, período marcado por centenas de mortos e

desaparecidos, são indispensáveis para compreender a história do país e evitar que momentos como esse se repitam. “É apenas com a preservação de dados sobre esse tema, sobre como a população considerou e apoiou a democracia ao final da ditadura militar e no início da democratização, [...] que será possível avaliar o sucesso da nossa construção democrática”, argumenta Rachel Meneguello, cientista política e pró-reitora de Pós-Graduação da Universidade Estadual de Campinas (Unicamp).

A pesquisadora enfatiza que os avanços recentes na preservação e no acesso a registros têm facilitado investigações em sua área e gerado resultados inovadores. “Há muitos acervos volumosos de dados documentais, dados oficiais agregados, dados de

natureza individual, que até anos atrás requeriam um esforço grande em processos de coleta e organização para viabilizar análises criativas”, lembra. “Não há comparação [de como os processos aconteciam nas décadas passadas] com o avanço que temos hoje. A integração de tecnologias digitais potencializa a coleta, interpretação, busca, leitura e análise de dados”, acrescenta.

Para Rachel Meneguello, o estudo de documentos históricos deve dialogar com informações inéditas, tanto em pesquisas quanto na formulação de políticas públicas. “A partir do conhecimento de dados históricos sobre uma população, uma oferta de serviços públicos ou carências específicas, se pode traçar tendências e padrões e revelar

necessidades ou lacunas que embasam políticas de governo”, explica.

O futuro da preservação e do acesso a dados históricos

Estudos sobre a ditadura brasileira, o “Caso Arquivado” e outros projetos que lidam com a digitalização de dados anteriores à popularização das tecnologias digitais enfrentam um cenário diferente do atual. “Quando observamos disputas de memória como essas, há um tempo entre o fato acontecer, a informação circular e aquele evento ser interpretado como história”, diz Thiago Nicodemo. “É uma grande disputa política de quem vai predominar, quem vai conseguir sobreviver ao tempo, contar a sua história”, complementa. “Hoje, esse tempo ficou encurtado a um instante. Essa disputa sobre qual narrativa vai predominar já vem quase embutida na evidência que circula. Se Anne Frank vivesse hoje, não ia dar tempo de escrever um diário; ia ser um post”, afirma.

O pesquisador prevê que as tecnologias modernas de informação facilitarão tentativas futuras de estudar e falar sobre o passado, mas também trarão novos questionamentos. “As novas mídias, as redes sociais e a facilidade que a gente tem com os dispositivos que produzem evidências,

como gravadores e celulares, permitem produzir cada vez mais evidências. Daí, nasce uma pergunta: como cuidar da informação que já nasce digital?”, indaga, questionando como a velocidade acelerada de registros, o acesso ampliado às informações e os dados produzidos no mundo virtual impactarão as narrativas de memória. “A gente vai ter mais ou menos memória?”, pergunta.

“Uma informação tem que ser guardada porque contém um direito, é testemunha de alguém que sofreu uma violência ou deve alguma coisa.”

Embora as respostas ainda sejam incertas, ele alerta para a necessidade de intensificar o investimento em infraestrutura para lidar com as novas demandas de armazenamento e acesso. “Normalmente, uma informação tem que ser guardada porque contém um direito, é testemunha de alguém que sofreu uma violência ou deve alguma coisa, mas estamos em um mundo que produz informações relevantes muito rápido e tem baixíssimas condições para guardá-las”, argumenta. Gilberto Lacerda dos Santos destaca que os

repositórios institucionais são alternativas indispensáveis: “são excelentes suportes para dados, a longo prazo”. Rodrigo Lima-Lopes concorda e ressalta que a padronização do acesso também é importante. “Repositórios organizados por instituições de pesquisa, como o Repositório de Dados da UNICAMP (REDU) e os dados de pesquisa política do Brazilian Political Corpus (BRPoliCorpus), sob minha coordenação, são exemplos interessantes”, lista.

Ainda que as conclusões do “Caso Arquivado” pressuponham lacunas decorrentes da destruição e perda de documentos, o uso de estratégias atuais de preservação e acesso a dados históricos, como a digitalização, oferece novos horizontes para a investigação. Ao recorrer ao resgate de informações históricas, pesquisas retrospectivas, como essa, podem não apenas revisitar, mas também reanalisar e reinterpretar narrativas históricas sob olhares inéditos e reconstruir partes importantes da história com uma precisão cada vez maior.

Bianca Bosso é especialista em Jornalismo Científico e Bacharela em Ciências Biológicas (Unicamp). Iniciou sua trajetória na Divulgação Científica no ano de 2018. Já desenvolveu pautas para revistas como Ciência & Cultura, ComCiência e Ciência Hoje, além de sites como Agência Bori, Jornal da Unicamp, Portal Campinas Inovadora e blog Ciência na Rua.