25°enancib
Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação
Informação, Decolonialidade e Direitos Difusos para o Desenvolvimento Sustentável
3 a 7 de novembro de 2025 - Rio de Janeiro

XXV ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – XXV ENANCIB

GT 8 – Dados, Informação e Tecnologia

O FORMATO DE ARQUIVO DIGITAL PDF/A: POSSIBILIDADES E LIMITAÇÕES DE USO

THE PDF/A DIGITAL FILE FORMAT: POSSIBILITIES AND LIMITATIONS OF USE

Dalton Garcia do Carmo – Universidade Federal de Minas Gerais (UFMG)
Cintia Aparecida Chagas – Universidade Federal de Minas Gerais (UFMG)

Modalidade: Resumo Expandido

Resumo: o formato PDF/A possui requisitos para a gestão e preservação de documentos digitais no longo prazo. O objetivo deste trabalho foi problematizar a utilização deste formato, apresentando os aspectos relacionados a padrões, níveis de conformidade, metadados, reconhecimento de caracteres, novas tecnologias digitais e padrões arquivísticos. A metodologia adotada foi a abordagem qualitativa e descritiva, e os métodos de pesquisa foram a pesquisa bibliográfica e documental. Sugeriu-se a continuidade da pesquisa com o foco na extração de dados, em metadados e no arquivamento da web. Concluiu-se que a padronização do formato foi importante para acessibilidade, confiabilidade e integridade do documento.

Palavras-chave: formato de arquivo digital PDF/A; preservação; metadados.

Abstract: the PDF/A format has requirements for the long-term management and preservation of digital documents. The objective of this study was to discuss the use of this format, presenting aspects related to standards, compliance levels, metadata, character recognition, new digital technologies, and archival standards. The methodology adopted was a qualitative and descriptive approach, and the research methods were bibliographic and documentary research. It was suggested that research should continue, focusing on data extraction, metadata, and web archiving. It was concluded that standardization of format was important for accessibility, reliability, and document integrity.

Keywords: PDF/A digital file format; preservation; metadata.

1 INTRODUÇÃO

O formato de arquivo digital PDF (*Portable Document Format* ou Formato de Documento Portátil) é considerado um formato “padrão de fato” e “padrão de direito”, em virtude de sua ampla utilização, reconhecimento por parte dos usuários, padronização e adoção em órgãos oficiais, segundo o Conselho Nacional de Arquivos (Conarq, 2013, p. 4).

O formato de arquivo digital PDF foi criado pela Adobe Systems no início da década de 1990 com o objetivo de troca de documentos digitais e se popularizou por representar os documentos digitais de maneira semelhante aos seus equivalentes em papel, além de sua universalidade e portabilidade em softwares e sistemas computacionais.

Em 2005, a Adobe Systems forneceu a especificação do formato PDF para a *International Organization for Standardization* (ISO), tornando-o um padrão aberto (PDF Association, 2025, p. 5). No mesmo ano, foi publicada a norma ISO 19005-1:2005 - Gerenciamento de documentos - Formato eletrônico de arquivo de documento para preservação de longo prazo, que especificou o formato de arquivo digital PDF no padrão PDF/A (que denominaremos neste trabalho por formato PDF/A). Existem outros padrões para o formato PDF, como, por exemplo, o PDF/X (artes gráficas e impressão profissional), o PDF/UA (para acessibilidade universal), o PDF/R (para documentos de imagens raster).

Normatizado pela ISO e pela Associação Brasileira de Normas Técnicas (ABNT), o PDF/A oferece uma variedade de recursos para representação de um documento nato-digital ou de um representante digital, importantes no contexto arquivístico. Devido à sua massiva utilização em diversos setores da sociedade e às especificidades de cada versão e nível de conformidade do padrão, pressupõe-se que alguns recursos desse formato são pouco aproveitados ou utilizados de maneira equivocada. Buscou-se, portanto, relacionar o formato PDF/A com conceitos arquivísticos e ferramentas digitais em evidência, tais como os metadados, o Reconhecimento Ótico de Caracteres (OCR), o *Records in Context* (RiC) e a Inteligência Artificial (IA), verificando as possibilidades e limitações de uso.

Os metadados são efetivos na gestão e preservação do formato PDF/A e devem ser descritos segundo o padrão da Norma Geral Internacional de Descrição Arquivística (ISAD(G)) ou o recém-criado padrão RiC. Por sua vez, a pesquisa, edição e recuperação da informação em um documento no formato PDF/A são aprimoradas com o uso da tecnologia OCR. Por último, as ferramentas de IA e OCR ajudam a reconhecer o texto de documentos, tornando-os pesquisáveis. Esse processo contribui para a gestão arquivística, para a confiabilidade e preservação dos documentos no formato PDF/A.

Tais observações justificam essa investigação, com o objetivo de caracterizar o formato PDF/A, os seus padrões e níveis de conformidade, as vantagens e as limitações, os possíveis impactos no tratamento dos documentos arquivísticos digitais, a manipulação e o uso da informação e dos dados frente às inovações digitais.

A necessidade de aprofundar a pesquisa foi constatada, concluindo-se que, apesar de o PDF/A ser fundamental para a gestão e preservação de documentos digitais, suas restrições técnicas limitam a extração e a utilização de dados e informações do seu conteúdo em ambientes web.

2 PROCEDIMENTOS METODOLÓGICOS

A presente pesquisa possui caráter exploratório e descritivo. Foi realizada por meio de pesquisa bibliográfica e documental, abrangendo artigos científicos, publicações, normas padronizadoras ISO e ABNT e relatórios técnicos, relevantes para a discussão. A coleta dos materiais para pesquisa foi concentrada na base de dados BRAPCI e no Portal de Periódicos da CAPES. Foram priorizados trabalhos publicados nos últimos cinco anos que relacionassem as características do formato PDF/A com metadados e inteligência artificial no contexto arquivístico. A análise do conteúdo permitiu selecionar e identificar categorias e significados relevantes à compreensão do tema investigado.

3 CARACTERÍSTICAS DO FORMATO DE ARQUIVO DIGITAL PDF/A E AS SUAS VERSÕES

Foco deste estudo, o formato PDF/A é orientado à paginação e tem por objetivo manter a aparência visual de um documento no ambiente digital ao longo do tempo, sem a dependência de ferramentas específicas e sistemas computacionais para a produção, utilização e armazenamento. Ele também é capaz de prover uma estrutura para registro do contexto de produção e uso em metadados no próprio documento, além de definir um esquema para a representação da estrutura lógica e de informações semânticas (ABNT, 2009).

Betsy Fanning (2017, p. 3) destaca que o formato PDF/A tem particularidades em relação a outros padrões PDF, que visam garantir a segurança e a integridade do documento. Ele proíbe a incorporação de áudio, vídeo, criptografia, aplicativos e códigos executáveis (.bin, .exe ou em *JavaScript)* no documento, eliminando o risco de infecção por programas perigosos, como vírus. O formato PDF/A determina que as fontes tipográficas e as informações de cores necessárias para a renderização estejam incorporadas no respectivo documento digital e a utilização do formato XMP para o armazenamento dos metadados que o identificam como um PDF/A.

3.1 Versões e níveis de conformidade do padrão PDF/A

O formato PDF/A possui quatro padrões normatizados pela ISO, conforme apresentado no Quadro 1.

Quadro 1 - Versões do padrão PDF/A e suas principais características

Versão do padrão PDF/A	**Características**
PDF/A-1 (ISO 19.005-1:2005) BASEADO NO PDF 1.4	Proíbe uma série de recursos prejudiciais à preservação digital, tais como: códigos executáveis, *JavaScript*, *hyperlinks* externos, inserção de áudio e vídeo. Garante a presença de metadados e de fontes embutidas (*embedding font*).
PDF/A-2 (ISO 19.005-2:2011) BASEADO NO PDF 1.7	Complementa a versão anterior, define um novo formato que considera as novas características decorrentes da evolução do formato PDF ou proibidas na 19005-1. Exemplo: o PDF/A-2 pode conter, como anexos, outros documentos no formato PDF/A. Foram permitidas as seguintes características: transparência, camadas, compressão JPEG2000 e assinatura digital.
PDF/A-3 (ISO 19.005-3:2012) BASEADO NO PDF 1.7	Acrescenta a possibilidade de incluir anexos em qualquer formato de arquivo digital, como .doc ou .xls, permitindo a reutilização e edição de documentos no fluxo de trabalho. Esses anexos, chamados "arquivos associados", armazenam dados ou arquivos fonte do PDF/A-3.
PDF/A-4 (ISO 19.005-4:2020) BASEADO NO PDF 2.0	Compatível com o padrão PDF/A-3. Tem foco na autodocumentação com metadados descritivos, administrativos e de proveniência. Aceita um nível de conformidade voltado a arquivos de Engenharia (PDF/E). Aceita *Javascript* sem a execução pelo visualizador (apenas armazena informações sobre a lógica de um formulário interativo).

Fonte: adaptado de Cristóvão, Batista e Rocha (2023, p. 7), FileFormat (2025)

Para todos os padrões e versões do formato PDF, existem os níveis de conformidade, que, segundo a PDF Association (2025), estão relacionados a um conjunto próprio de regras e requisitos. Os níveis de conformidade do padrão PDF/A são caracterizados assim:

Quadro 2 - Níveis de conformidade para o padrão PDF/A

Níveis	**Características da conformidade**
a - ACESSÍVEL (*ACCESSIBLE*)	Atende a todos os requisitos para o padrão, incluindo a estrutura lógica do documento e sua ordem correta de leitura. O texto deve ser extraível (OCR em imagens) e a estrutura lógica deve corresponder à ordem natural de leitura. As fontes usadas devem atender a requisitos rigorosos. Aplicável aos padrões: PDF/A-1, PDF/A-2 e PDF/A-3.
b - BÁSICO (*BASIC*)	Garante que o conteúdo do documento seja reproduzido sem ambiguidade. Os arquivos de nível B são mais fáceis de criar do que os de nível A, mas o nível B não garante 100% de extração de texto ou capacidade de pesquisa. Isso não significa necessariamente que o conteúdo pode ser reutilizado sem problemas. Aplicável aos padrões: PDF/A-1, PDF/A-2 e PDF/A-3.
u – *UNICODE*	Foi introduzido junto com o PDF/A-2. Ele expande o nível B de conformidade para especificar que todo texto pode ser mapeado para códigos de caracteres *Unicode* padrão. Aplicável aos padrões: PDF/A-2 e PDF/A-3.
f - ARQUIVOS (*FILES*)	Este nível de conformidade permite a incorporação de anexos em qualquer formato digital e atua como um sucessor do padrão PDF/A-3. Aplicável ao padrão PDF/A-4.

e - ENGENHARIA (*ENGINEERING*)	Sucessor do padrão PDF/E, direcionado à incorporação de anexos comuns da área da Engenharia em 3D e arquivos *Rich Media*. Aceita anexos de quaisquer formatos, sendo capaz de processar ações em *JavaScript* se solicitado pelo usuário. Aplicável ao padrão PDF/A-4.

Fonte: adaptado de Cristóvão, Batista e Rocha (2023) e PDF Association (2025)

Cristóvão, Batista e Rocha (2023, p. 3) ressaltam que os arquivos em PDF/A também podem ser categorizados como:

Quadro 3 – Categorias do PDF/A relacionadas à inserção de metadados

Categorias	Descrição
AUTOCONTIDO OU AUTOSSUFICIENTE	Possui todas as informações necessárias para a sua exibição e é qualificado como PDF/A-3.
AUTODOCUMENTADO	Possui metadados semânticos incorporados, isto é, metadados que o descrevem, tais como título, autores, data de criação etc.
AUTORREFERENCIADO	Possui metadados de identificação única incorporados, tais como *Digital Object Identifier* (DOI), *International Standard Serial Number* (ISSN), ou *International Standard Book Number* (ISBN), permitindo que as máquinas de busca recuperem integralmente o conjunto de metadados por meio de uma dessas identificações únicas. Essa categoria é um subconjunto da categoria dos autodocumentados.

Fonte: adaptado de Cristóvão, Batista e Rocha (2023)

Salienta-se que o formato de arquivo digital PDF/A é um tipo de imagem vetorial, que captura e representa o documento original com a mesma aparência e estrutura, possibilitando a ampliação da imagem sem a perda de resolução e ocupando menos espaço de armazenamento. Diferente de outros formatos de arquivo de imagem, como o JPEG, PNG ou TIFF, que são do tipo raster (*bitmaps*), comumente recomendáveis para o armazenamento e preservação de imagens e fotos, e que permitem a edição de cada *pixel* de uma imagem.

Por fim, destaca-se a norma ISO 32000-1:2008, que, segundo a PDF Association (2025), diz respeito ao gerenciamento de documentos em PDF e se destina aos desenvolvedores de softwares para criação, gravação, leitura, edição e interação com o formato PDF/A.

3.2 Reconhecimento ótico de caracteres

O reconhecimento de caracteres contribui para a extração de dados, para a indexação automática do documento e conversão do conteúdo em outros formatos de arquivo digital. São exemplos de tecnologias de reconhecimento de caracteres (Hyperscience, 2020):

- OCR (*Optical Character Recognition*): O reconhecimento ótico de caracteres permite a conversão de caracteres gerados mecanicamente, impressos ou datilografados, e para conversão de documentos impressos em texto.

- ICR (*Intelligent Character Recognition*): O reconhecimento inteligente de caracteres possibilita a conversão de caracteres gerados de forma manuscrita, porém é menos preciso se considerarmos as variações na leitura de caligrafia.
- IDP (*Intelligent Document Processing*): O processamento inteligente de documentos é uma tecnologia de automatização de fluxos de trabalho que digitaliza, lê, extrai, categoriza e organiza informações significativas em formatos acessíveis.

Mesmo com os avanços tecnológicos, as ferramentas de reconhecimento ótico de caracteres não são totalmente precisas. Gil-Leiva *et al*. (2022, p. 2) afirmam que a variação na estrutura e no leiaute do documento digital no formato PDF dificultam o seu processamento e a obtenção de informações de título, autores, quadros e imagens, tornando essa tarefa complexa mesmo quando realizada por softwares específicos.

A qualidade e a resolução da imagem digitalizada, a caligrafia no documento e os modelos de treinamento para reconhecimento de caracteres utilizados por IA (aprendizado de máquina) influenciam diretamente nos resultados. A análise de leiaute do documento digital é desafiadora, pois envolve a correta segmentação da página e a manutenção da ordem de leitura. Nesse aspecto, a IA tem demonstrado potencial para melhorar a qualidade dos textos processados por OCR, reduzindo erros e aumentando a precisão na identificação de conteúdo (Kavčič Čolić; Hari, 2024, p. 190).

Ressalta-se que o OCR não é um recurso intrínseco ao formato PDF, e sim uma ferramenta complementar. No caso de representantes digitais (documentos digitalizados), o OCR é um recurso essencial para preservação e acesso ao documento digital.

3.3 O protocolo XMP

A partir da versão 1.4 do formato de arquivo PDF, foi criado um mecanismo capaz de incorporar metadados, denominado de fluxo de metadados, que se utiliza de uma estrutura chamada de *Extensible Metadata Platform* (XMP), normatizada pela ISO 16684-2019 (Cristóvão; Batista; Rocha, 2023). Ele fornece informações sobre o arquivo, como identificação, descrição e aspectos técnicos especificados em formato XMP e incorporados ao próprio PDF/A (ABNT, 2009). Para fins de ilustração, o quadro 4 apresenta os metadados do documento (dicionário de informações do documento) e a codificação correspondente dos metadados XMP (PDF Association, 2025):

Quadro 4 – Metadados XMP em um PDF/A-1

Dicionário de informação do documento		**Codificações XMP**	
ENTRADA	**TIPO PDF**	**PROPRIEDADE**	**TIPO XMP**
Title	texto string	dc:title	Lang Alt
Author	texto string	dc:creator	seq ProperName
Subject	texto string	dc:description["x-default"]	Texto
Keywords	texto string	pdf:Keywords	Texto
Creator	texto string	xmp:CreatorTool	AgentName
Producer	texto string	pdf:Producer	AgentName
CreationDate	data	xmp:CreateDate	Data
ModDate	data	xmp:ModifyDate	Data

Fonte: Adaptado de Technote 0003 – Metadata in PDF/A-1 (PDF Association, 2025)

Destaca-se que, segundo Betsy Fanning (2017, p. 11), o "XMP é usado para codificar os metadados do documento dentro do arquivo PDF/A e é baseado no *Resource Description Framework* (RDF) (ou Estrutura de Descrição de Recursos)". Ainda segundo a autora, o "RDF é o alicerce da web semântica e, usando o XMP, os sistemas e aplicativos podem acessar e compreender os metadados dos documentos que manipulam".

A incorporação de metadados no próprio documento no formato PDF/A contribui para a transparência em relação à produção, edição e propriedade intelectual, e favorece a interoperabilidade entre sistemas computacionais, diminuindo a necessidade de edição e revisão manual. A Adobe oferece um kit de desenvolvimento de software (SDK) denominado XMP *Toolkit* para inserção ou edição de metadados em documentos digitais no formato PDF.

3.4 Limitações e desafios para o PDF/A

O formato PDF/A não deve ser apontado como a única solução de preservação. Segundo Betsy Fanning (2017, p. 20), ele precisa estar integrado a outros componentes e procedimentos estabelecidos em uma estratégia de preservação ampla e consistente com a infraestrutura de preservação, como *backups*, verificações de integridade e documentação.

Por exemplo, a migração de um documento PDF para PDF/A nem sempre é recomendável. Se a fonte de texto não estiver incorporada no PDF de origem, o novo documento PDF/A precisará inserir uma nova fonte para não comprometer a sua visualização. Algumas fontes de texto possuem direitos autorais (Fanning, 2017, p. 13).

Em relação às assinaturas eletrônicas, a migração de formato PDF para PDF/A quebrará as assinaturas existentes, não sendo possível recuperá-las (Fanning, 2017, p. 13). Além disso, assinaturas digitais são permitidas a partir do padrão de formato PDF/A-2 e devem ser compatíveis com o padrão PDF *Advanced Electronic Signatures* (PAdES).

A constante evolução das ameaças de vírus para computador e a complexidade estrutural do formato PDF tornam a análise de *malwares* uma atividade problemática. Singh, Tapaswi e Gupta (2020) destacam a vulnerabilidade em programas leitores de PDF, como o Adobe Acrobat Reader e o Foxit Reader, que regularmente são alvos de ataques. A busca e análise dessas ameaças geram um alto custo computacional e resultados ambíguos. Nesse quesito, o formato PDF/A e as suas restrições garantem mais segurança ao documento.

3.5 A utilização do PDF/A no ambiente digital

A natureza estática do formato PDF/A torna-o menos eficiente em ambientes que requerem interoperabilidade de dados e conformidade semântica, ainda que seja possível realizar a extração de seu conteúdo. De fato, o formato PDF/A foi projetado para captura de imagem, visualização, impressão de documentos em meio digital e não para apresentar relações semânticas entre dados (Fanning, 2017, p. 12).

A dificuldade na extração de dados e de identificação de informações estruturadas, sejam estes dados, metadados e paradados, dificulta a integração do padrão PDF/A com os modelos RiC, com o padrão RDF e o *Linked Open Data* (LOD) (dados abertos vinculados), temas que estão em evidência no contexto da Ciência da Informação (CI).

Segundo Feliciati e Duranti (2025), enquanto os metadados dizem respeito aos dados estruturados e formalizados relativos aos documentos arquivísticos, os paradados são fundamentais para retratar o contexto de produção e de gestão dos documentos, com informações mais amplas e menos detalhadas do que os metadados. Em sua pesquisa, as autoras destacam que os paradados asseguram a garantia, a responsabilização e a transparência no uso de sistemas e ferramentas de IA em ambientes arquivísticos.

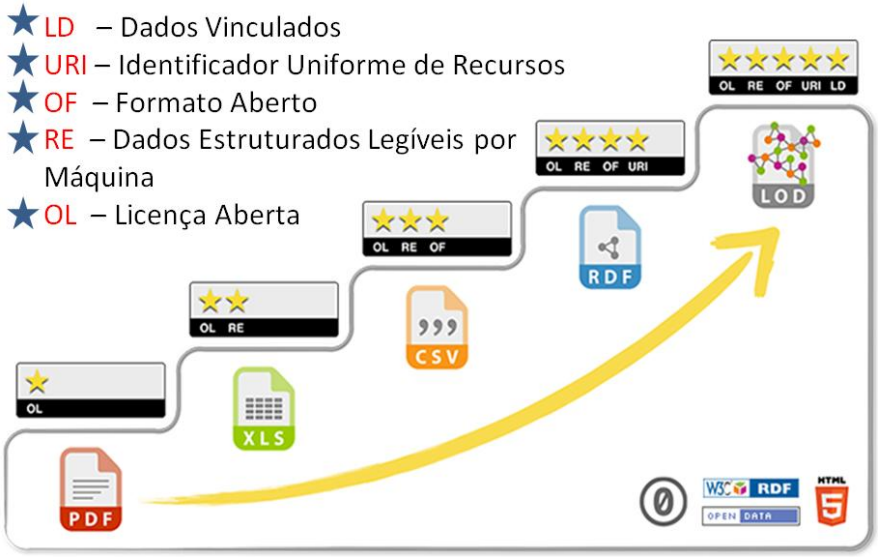
Em relação ao modelo RiC, Borges e Roncaglio (2024) afirmam que ele surgiu como uma resposta à evolução das tecnologias de informação e comunicação (TICs), o avanço da web semântica e dos dados abertos, marcando a transição da "web de documentos" para a "web de dados". A “web de dados”, por sua vez, compreende a aplicação da inteligência

artificial para organizar e recuperar dados (Borges; Roncaglio, 2024, p. 14), transcendendo a busca "por páginas" em documentos no formato PDF/A.

A extração automática de dados, como "data", "autor" e "palavras-chave" no conteúdo do documento a partir das ferramentas de OCR, pode tornar o processo de representação do conteúdo na forma de LOD complexo e sujeito a erros. Nesse quesito, as buscas por uma informação no documento e por temas relacionados em outros documentos são menos precisas, prejudicando a integração em sistemas de gestão de conteúdo, a criação de conexões entre documentos e a construção de bases de conhecimento mais robustas.

Tim Berners-Lee, inventor da web, sugeriu um esquema de estrelas para qualificar as publicações de dados abertos e facilidade de utilização pelas pessoas:

Figura 1 – As 5 Estrelas dos Dados Abertos



Fonte: adaptado de 5stardata (2012)

Nesse esquema, o formato PDF possui uma estrela e se destaca apenas por possuir licença aberta. Com cinco estrelas, estão os LOD que possibilitam ao usuário manipular, criar referências, reutilizar e descobrir dados relacionados ao conteúdo que pesquisou na web. Nesse aspecto, torna-se relevante identificar como os documentos no formato PDF/A serão incorporados à "web de dados", com a utilização de ferramentas de IA, sem que haja perda informacional e garantindo a preservação dos documentos no longo prazo.

4 RESULTADOS PARCIAIS

A continuidade dessa pesquisa pode seguir alguns caminhos. Conforme sinaliza Betsy Fanning (2017, p. 11), o aproveitamento de metadados XMP incorporados ao documento digital no formato PDF/A contribui para o seu arquivamento, para a recuperação e eficácia nas

buscas. A utilização de um vocabulário controlado padronizado torna-se útil para a descrição correta e interoperação de documentos e sistemas. Além disso, explorar os aspectos relacionados a dados, metadados e paradados é preponderante para a preservação do documento digital.

Quanto ao reconhecimento de caracteres, em conjunto com a inteligência artificial, essa ferramenta pode possibilitar a integração do formato de arquivo PDF/A ao padrão RiC, ao padrão RDF e LOD. No contexto da web, é importante verificar as interlocuções entre o formato de arquivo PDF/A e o formato de arquivo WARC (*Web Archive*), projetado para arquivar conteúdo da *web* (ISO 28500). Destaca-se também o formato de arquivo EA-PDF (*Archival Email Format based on PDF/A*), desenvolvido para a preservação de e-mail.

Mediante os recursos oferecidos e os riscos inerentes ao ambiente digital, em prol de uma efetiva gestão, preservação e consonância com as normas existentes (inclusive a legislação brasileira), auditar e certificar a produção de documentos em formato PDF/A são atividades que favorecem o acesso e a preservação da informação. As políticas arquivísticas, os procedimentos relacionados à gestão de documentos e os seus instrumentos são essenciais para orientar e qualificar o uso do formato de arquivo digital PDF/A.

5 CONSIDERAÇÕES FINAIS

A padronização do formato PDF/A por normas ISO e ABNT mostrou-se fundamental para a preservação do documento digital, mantendo a sua acessibilidade, confiabilidade e integridade. Constatou-se a existência de uma vasta documentação técnica pertinente ao formato PDF/A. Restrições relacionadas à incorporação de recursos multimídia, fontes, reconhecimento de texto, metadados, se destacam quanto à utilização do formato PDF/A, protegendo-o de ameaças. No entanto, dificultam a extração de informações e a preservação de conteúdo interativo.

A distinção entre padrões e níveis de conformidade pode interferir na gestão dos documentos digitais, considerando as especificidades de cada um deles. Sugere-se que novos estudos insiram o formato PDF/A sob a perspectiva da CI e da Arquivologia, aprimorando a produção, recuperação, acesso e preservação de documentos digitais no formato PDF/A, o uso dos metadados, a acessibilidade, a propriedade intelectual e a interação com IA e demais ferramentas tecnológicas.

REFERÊNCIAS

5STARDATA. 5 **Estrelas dos Dados Abertos**. 2012. Disponível em: https://5stardata.info/pt-BR/. Acesso em: 29 abr. 2025.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS (ABNT). A**BNT NBR ISO 19005:** Gerenciamento de documentos - Formato eletrônico de arquivo de documento para preservação de longo prazo. Parte 1: Uso do PDF 1.4 (PDF/A-1). Rio de Janeiro: ABNT, 2009. 35 p.

BORGES, Thiago Almeida Rodrigues; RONCAGLIO, Cynthia. Renovação teórica e inovação tecnológica em descrição de arquivos: o modelo conceitual Records in Contexts (RiC). **Acervo**, Rio de Janeiro, v. 37, n. 3, p. 1-29, set./dez. 2024. Disponível em: https://revista.arquivonacional.gov.br/index.php/revistaacervo/article/view/2135. Acesso em: 13 ago. 2025.

CONSELHO NACIONAL DE ARQUIVOS (CONARQ). **Resolução nº 38, de 09 de julho de 2013.** Dispõe sobre a adoção das Diretrizes do Produtor - A Elaboração e a Manutenção de Materiais Digitais: Diretrizes Para Indivíduos e Diretrizes do Preservador - A Preservação de Documentos Arquivísticos digitais: Diretrizes para Organizações, 2013. Disponível em: https://www.gov.br/conarq/pt-br/centrais-de-conteudo/publicacoes/conarq_diretrizes_produtor_preservador_resolucao_38.pdf. Acesso em: 29 abr. 2025.

CRISTOVÃO, Henrique Monteiro; BATISTA, Willian Alves; ROCHA, Bruna Morêto Sibaldo. Documentos digitais em formato PDF autocontidos, autorreferenciados e autodocumentados como suporte à publicação ampliada. **ÁGORA**: Arquivologia em debate, [S. l.], v. 33, n. 67, p. 1–25, 2023. Disponível em: https://agora.emnuvens.com.br/ra/article/view/1151. Acesso em: 29 abr. 2025.

FANNING, Betsy. **Preservation with PDF/A**. 2017. Disponível em: https://www.dpconline.org/docs/technology-watch-reports/1707-twr17-01-revised/file. Acesso em: 29 abr. 2025.

FELICIATI, Pierluigi; DURANTI, Luciana. The responsible use of Artificial Intelligence in archives through the use of paradata. **JLIS.it** , *[s. l.]*, v. 16, n. 2, p. 1–9, 2025. Disponível em: https://www.jlis.it/index.php/jlis/article/view/636. Acesso em: 13 ago. 2025.

FILEFORMAT. **What is a PDF file?** 2025. Disponível em: https://docs.fileformat.com/pdf/. Acesso em: 29 abr. 2025.

GIL-LEIVA, Isidoro; FUJITA, Mariângela Spotti Lopes; REDIGOLO, Franciele Marques; SARAN, Jordan Ferreira. Extracción de información de documentos PDF para su uso en la indización automática de e-books. T**ransinformação**, v. 34, e210069, 2022. Disponível em: https://doi.org/10.1590/2318-0889202234e210069. Acesso em: 29 abr. 2025.

HYPERSCIENSE. **Acronyms Explained: IDP vs. OCR & ICR**. 2020. Disponível em: https://www.hyperscience.ai/blog/acronyms-explained-idp-vs-ocr-icr/. Acesso em: 29 abr. 2025.

KAVČIČ ČOLIĆ, Alenka; HARI, Andreja. Improving accessibility of digitization outputs: EODOPEN project research findings. **Digital Library Perspectives**, [*s. l.*], v. 40, n. 2, p. 187–211, 2024. Disponível em: DOI: 10.1108/DLP-09-2023-0080. Acesso em: 29 abr. 2025.

PDF ASSOCIATION. **PDF Standards**. 2025. Disponível em: https://pdfa.org/pdf standards/. Acesso em: 29 abr. 2025.

SINGH, Priyansh; TAPASWI, Shashikala; GUPTA, Sanchit. Malware Detection in PDF and Office Documents: A survey. **Information Security Journal**: A Global Perspective, [*s. l.*], p. 1-28, 13 fev. 2020. Disponível em: DOI: 10.1080/19393555.2020.1723747. Acesso em: 29 abr. 2025.

TECHNOTE 0003. **Metadata in PDF/A-1**. PDF Association. 2008. Disponível em: https://pdfa.org/resource/technical-note-tn0003-metadata-in-pdfa-1/. Acesso em: 29 abr. 2025.