Documentos digitais em formato PDF autocontidos, autorreferenciados e autodocumentados como suporte à publicação ampliada

Henrique Cristovão Universidade Federal do Espírito Santo, ES, Brasil
https://orcid.org/0000-0003-2011-7022
hmcristovao@gmail.com

Willian Alves Batista Universidade Federal do Espírito Santo, ES, Brasil
https://orcid.org/0000-0002-5925-3817
willian.alves.b15@gmail.com

Bruna Morêto Sibaldo Rocha Universidade Federal do Espírito Santo, ES, Brasil
https://orcid.org/0000-0002-0987-6768
bruna.m.rocha@edu.ufes.br

Resumo O contexto e o uso de documentos arquivísticos mudam ao longo do tempo tornando-os agentes ativos numa sociedade com novas dinâmicas. Documentos digitais em formato PDF, sejam autocontidos, autorreferenciados ou autodocumentados, são requeridos em várias áreas, como na publicação ampliada. Objetivos: investigar os elementos conceituais e as ferramentas computacionais apropriadas para criação de documentos digitais em formato PDF autocontidos, autorreferenciados e autodocumentados, atendendo aos princípios FAIR e em um contexto de publicação ampliada; e criar um fluxo procedimental para adequar os documentos digitais em formato PDF, oriundos de trabalhos de conclusão de curso, a tornarem-se autocontidos e autodocumentados. Métodos: pesquisa exploratória descritiva e de levantamento bibliográfico. O cenário para a realização do experimento foi um acervo de trabalhos de conclusão de curso. Resultados: apresentação conceitual das caracterizações de documento autocontido, autorreferenciado e autodocumentado e sua pertinência com os princípios FAIR e com a publicação ampliada; escolha de três ferramentas para incorporação de anexos e metadados e validação para o padrão PDF/A, baseada em 12 funções elaboradas considerando-se aspectos relevantes ao contexto da pesquisa; fluxo procedimental com demonstração de uso para orientar o usuário no depósito do trabalho de conclusão de curso. Conclusões: documentos com anexos e metadados incorporados são importantes como suporte à publicação ampliada devido a vinculação do seu conteúdo aos seus dados complementares.

Palavras-chave Documento digital em formato PDF. Documento autocontido. Documento autorreferenciado. Documento autodocumentado. Publicação ampliada.

Self-contained, self-referenced and self-documented digital documents in PDF format supporting enhanced publication

Abstract The context and use of archival documents change over time, making them active agents in a society with new dynamics. Digital documents in PDF format, whether self-contained, self-referenced, or self-documented, are required in many areas, such as enhanced publication. Objectives: to investigate the conceptual elements and the appropriate computational tools for creating self-contained, self-referenced and self-documented to digital documents in PDF format, meeting FAIR principles and in a context of enhanced publication; and create a procedural flow to adapt digital documents in PDF format, originating from course conclusion paper, to become self-contained and self-documented. Methods: descriptive exploratory research and bibliographic survey. The scenario for carrying out the experiment was a collection of course conclusion paper. Results: conceptual presentation of the characterizations of self-contained, self-referenced and self-documented documents and their relevance to the FAIR principles and extended publication; choice of three tools for the incorporation

of attachments and metadata and PDF/A validation, based on 12 functions elaborated on aspects relevant to the research context; procedural flow with demonstration of use to guide the user in the course conclusion paper deposit. Conclusions: documents with attachments and embedded metadata are important as support for enhanced publication due to the linking of their content to their complementary data.

Keywords Digital documents in PDF format. Self-contained document. Self-referenced document. Self-documented document. Enhanced publication.

Submetido em 02/09/2022
Aprovado em 05/06/2023
Publicado em 06/07/2023

1 INTRODUÇÃO

O formato[1] de documentos digitais PDF (*portable document format* - formato de documento portável) é universal, popular, com alto nível de portabilidade em *hardware*, *software* e sistemas operacionais. Orientado a páginas, possui foco na preservação da aparência para o usuário, desde a produção até o arquivamento e a sua utilização. Contudo, esse formato sempre foi considerado ruim do ponto de vista semântico para interpretação automática por máquina devido a sua própria natureza de concepção, pois numa ótica orientada a dados, ele tem baixa granularidade, isto é, não oferece acesso detalhado ou categorizado de forma explícita às informações, as quais ele representa. Por outro lado, esse formato evoluiu nos últimos anos para suportar a adição de anexos com dados e a incorporação de metadados semânticos aproximando-se do movimento contemporâneo *data driven*[2]. Dessa forma, abriu-se um leque de possibilidades para que os documentos fossem interpretados por máquinas usando-se *softwares*, tais como, gerenciadores bibliográficos, motores de busca, ferramentas para descoberta de conhecimento, plataformas arquivísticas etc.

Outra vantagem na incorporação de dados, metadados semânticos e anexos em um documento PDF é que eles ficam acoplados ao seu conteúdo, em um único arquivo digital, diminuindo as possibilidades de perda por separação involuntária devido à falta de organização nas bases de dados ou por descuido do profissional gestor dos documentos. Cook (2012) chama

[1] O termo 'formato' aqui empregado ao documento digital vai ao encontro da terminologia usada por Rondinelli (2013) em seu livro 'O documento arquivístico ante a realidade digital uma revisão conceitual necessária'. Rondinelli destaca a diferença entre forma e formato. Assim, 'formato de arquivo' é a "[...] especificação de regras e padrões descritos formalmente para interpretação dos bits constituintes de um arquivo digital" (CÂMARA TÉCNICA DE DOCUMENTOS ELETRÔNICOS, 2020).

[2] *Data driven* é um movimento amplo em várias áreas do conhecimento que direciona as atividades para que sejam feitas prioritariamente baseada em dados e, dessa forma, construírem melhores soluções para diversos problemas.

atenção para novas tendências em que o contexto e o uso dos documentos arquivísticos mudam ao longo do tempo, uma vez que eles não se comportam mais como objetos passivos, ou simples registros de evidências, mas agentes ativos que desempenham papéis permanentes na vida de indivíduos, organizações e sociedade, devendo ser continuamente renovados para acompanhar as novas dinâmicas dos documentos.

No contexto da comunicação científica, o conceito de publicação ampliada refere-se a uma publicação que agrega, de forma padronizada, o texto do trabalho, os conjuntos de dados e os metadados (SALES; SOUZA; SAYÃO, 2014). A incorporação de elementos complementares no documento do texto principal, tais como: tabelas, algoritmos, imagens, áudios e vídeos, podem facilitar o entendimento da pesquisa, ao qual o documento se refere, como "[...] também podem ser aproveitados em outros contextos, como na colaboração entre pesquisadores, de onde emerge a necessidade de ambientes que propiciem a ampliação do espectro de documentos disponíveis destas publicações" (RODRIGUES; SANT'ANA, 2016, p. 5).

Um documento digital em formato PDF que possui todas as informações necessárias para a sua exibição e utilização é denominado de **autocontido**, ou autosuficiente, e é qualificado como PDF/A[3]. Em consonância com os princípios da publicação ampliada, o padrão PDF/A permite a adição de arquivos de diversos formatos possibilitando a guarda de dados importantes que devem ficar vinculados ao documento base, tais como: dados oriundos de documentos administrativos, diagramas de engenharia, dados de pesquisas científicas etc.

Um documento PDF com metadados de identificação única incorporados, tais como DOI[4], ISSN[5], ou ISBN[6] é denominado de **autoreferenciado**, pois permite que as máquinas de busca recuperem integralmente o conjunto de metadados que o descrevem por meio de uma dessas identificações únicas.

Um documento PDF com metadados semânticos incorporados, isto é, metadados que o descrevem tais como título, autores, data de criação etc. é denominado de **autodocumentado**.

[3] PDF/A é uma família de padrões ISO para formas restritas de PDF destinado a ser adequado para preservação a longo prazo de documentos digitais orientados a página (LOC, 2020a; SAA, 2005a).
[4] DOI (digital object identifier) fornece uma infraestrutura técnica e social para o registro e uso de identificadores interoperáveis persistentes para uso em redes digitais e segue o padrão ISO 26324. Disponível em: https://www.doi.org/.
[5] ISSN (international standard serial number) é um número internacional de identificação única para publicações seriadas. Disponível em: https://www.issn.org/.
[6] ISBN: (international standard book number) é um número internacional de identificação única para livros. Disponível em: https://isbn.org/

Dessa forma, a categoria dos documentos autorreferenciados é um subconjunto da categoria dos autodocumentados.

Tanto os documentos PDF autorreferenciados quanto os documentos PDF autodocumentados auxiliam os processos de reconhecimento, interpretação e cadastro de metadados em *softwares* como os gerenciadores bibliográficos e gerenciadores de repositórios e, portanto, auxiliam o processo da publicação ampliada. Também podem ser utilizados por *softwares* específicos para a Arquivologia, como os ambientes de gestão de documentos, ambientes de preservação e ambientes de acesso e difusão.

Os documentos, independente do seu suporte e de suas informações, necessitam perdurar conforme avaliação, pois representam um tipo de conhecimento que foi gerado ou recebido no curso de atividades pessoais ou institucionais com finalidade de oferecer estimativas e conclusões relativas às essas atividades (DURANTI, 1994). Nesse contexto, a recuperação de informações contidas em documentos sempre foi algo desafiador cujo debate tem sido amplificado por movimentos como os princípios FAIR[7], que enfatizam a importância de aperfeiçoar a forma como as máquinas conseguem usar os dados de forma automática e a sua reutilização por indivíduos (WILKINSON et al., 2016) favorecendo o processo de reprodutibilidade e continuidade de pesquisas na ciência, como a publicação ampliada. Rodrigues e Sant'ana (2016) alertam que há carência de ambientes que propiciem a efetivação da publicação ampliada.

Dessa forma, o problema de pesquisa do presente trabalho é como criar ou adaptar documentos digitais em formato PDF para atenderem aos princípios FAIR em um contexto de publicação ampliada? Para viabilizar as experimentações que foram realizadas na presente pesquisa, foi utilizado o acervo de trabalhos de conclusão do curso (TCC) de graduação em Arquivologia da Universidade Federal do Espírito Santo.

Os objetivos do trabalho são: (i) investigar os elementos conceituais e as ferramentas computacionais apropriadas para criação de documentos digitais em formato PDF autocontidos, autorreferenciados e autodocumentados, atendendo aos princípios FAIR e em um contexto de publicação ampliada; (ii) criar um fluxo procedimental para adequar os documentos digitais em formato PDF, oriundos de TCCs, para tornarem-se autocontidos e autodocumentados, usando as ferramentas computacionais selecionadas.

[7] FAIR é um acrônimo para findable, accessible, interoperable e reusable. Refere-se a um movimento mundial para orientar a construção de bases de dados e documentos com intuito de aumentar a capacidade de serem encontrados, acessíveis, interoperáveis e reutilizáveis.

O artigo apresenta, na seção 2, os fundamentos de documentos digitais em formato PDF, tanto os autocontidos PDF/A com os seus benefícios e riscos de uso, quanto os autorreferenciados e autodocumentados por meio da incorporação de metadados. A seção 3 aborda os procedimentos metodológicos. A seção 4 apresenta os resultados apontando as ferramentas selecionadas e um fluxo procedimental com exemplo completo de sua execução. A seção 5 faz as considerações finais.

2 DOCUMENTOS DIGITAIS EM FORMATO PDF NA PUBLICAÇÃO AMPLIADA

Um documento é uma unidade de registro de informações, qualquer que seja o formato ou o suporte, sendo que o documento digital diferencia-se pelo fato da informação ser codificada em dígitos binários cuja informação é acessível e interpretável por meio de um sistema computacional (CONARQ, 2020). Contudo, o suporte digital provocou mudanças significativas na forma de conceber, organizar e usar os documentos. Especificamente o documento digital em formato PDF possui características de portabilidade que garantem que sua criação e exibição ocorra por grande variedade de *softwares* e plataformas, de modo que a forma de exibição de suas informações textuais ou multimidiáticas seja preservado. Apesar disso, o PDF não é totalmente independente de plataforma, pois requer um leitor para renderizar o arquivo, embora o *software* de leitura exista para a maioria das plataformas (SAA, 2005a).

2.1 O FORMATO BASE PDF

Em 1991, por meio do projeto intitulado “The Camelot Project”, a Adobe[8] desenvolveu um formato para a preservação da aparência, de documentos advindos de qualquer proveniência, que viria a se tornar, no ano seguinte, o formato denominado de PDF (ADOBE ACROBAT, 2022). Mas, a primeira versão, de fato, foi lançada em 1993, ainda pela Adobe, e continuando a ser um formato proprietário até 2008, quando a versão 1.7 foi lançada e normalizada por um comitê da International Organization of Standardization (ISO) por meio da norma ISO 32000-1:2008[9]. Desde então, tornou-se um padrão livre com sua especificação controlada pela ISO que ficou responsável pelas versões futuras (WHEATLEY, 2019). A versão mais recente é a 2.0, lançada em 2017, e

[8] Adobe Inc. é uma multinacional americana que desenvolve programas de computador com sede em San Jose, Califórnia. Disponível em: https://www.adobe.com/.

[9] Padrão ISO 32000 especifica um formato digital para a implementação de gestão de documentos possibilitando o uso e a troca de documentos eletrônicos, independentemente do ambiente tecnológico. Disponível em: https://www.iso.org/standard/51502.html/.

padronizada pela ISO 32000-2:2017. Wheatley destaca que a versão 2.0 não foi criada com a pretensão de substituir a versão 1.7, mas para aprimorá-la e oferecer mais uma opção de uso do formato de arquivo PDF.

Em ambas versões, 1.7 e 2.0, o formato de arquivo PDF, juntamente com o *software* para criar, visualizar, imprimir e processar, deve atender um conjunto de requisitos (LOC, 2020d): (i) preservar a fidelidade do documento independente do dispositivo, plataforma e *software*; (ii) permitir a fusão de diversas fontes e manter a integridade de todos os documentos originais; (iii) disponibilizar um modelo de metadados extensível no nível de documento e objeto; (iv) oferecer edição colaborativa de documentos de vários locais ou plataformas; (v) permitir assinaturas digitais para certificar a autenticidade; (vi) oferecer segurança e permissões para que o criador mantenha o controle do documento e direitos associados; (vii) dar acessibilidade de conteúdo para pessoas com deficiência; (viii) permitir a extração e reutilização de conteúdo para uso com outros formatos de arquivo e aplicativos; e (ix) oferecer formulários eletrônicos para coletar e/ou representar dados dentro de sistemas diversos.

Apesar do formato base PDF conseguir atender à maioria das demandas, Wheatley (2019) observa que existem vários problemas, como arquivos inválidos ou mal formados em coleções depositadas e o impacto negativo disso na preservação de documentos. O autor também alerta para problemas associados à falta de validação do formato de arquivo PDF bem como a riscos relacionados a fontes inexistentes na exibição do seu conteúdo. Uma variação do formato PDF base, denominada de padrão PDF/A, veio para eliminar esses e outros problemas, principalmente relacionados ao arquivamento digital. O presente artigo tem foco nesse padrão PDF/A, que será tratado nas próximas subseções.

Além do PDF/A, existem outros padrões para o formato de arquivo PDF (OETTLER, 2013): PDF/X com foco em ações de impressão para conservação de fontes, imagens e cores; PDF/VT que possui suporte a impressão de dados variáveis como anúncios personalizados, faturas etc.; PDF/UA com foco em acesso universal, voltado especialmente para pessoas com deficiência; e PDF/E com foco em documentos de engenharia contendo principalmente diagramas representados por imagens vetoriais.

2.2 O DOCUMENTO AUTOCONTIDO PDF/A E SEUS PADRÕES

O padrão PDF/A (*portable document format/archive* - formato de documento portável para arquivamento) tem como principal característica a reprodutibilidade ao longo do tempo. Ele é

100% autocontido, isto é, possui a incorporação de todas as informações necessárias para sua exibição (VASILESCU, 2009). Isso inclui, entre outras coisas, o conteúdo propriamente dito (texto, imagens e gráficos vetoriais), as fontes utilizadas e as informações de cor, não sendo permitidas dependências a fontes externas (ABNT NBR ISO, 2019).

Amplamente usado no mundo e obrigatório em várias instituições, o formato de arquivo PDF/A é recomendado como padrão de preservação pelo Arquivo Nacional (2016) e também pelo Decreto Nº 10.278 para documentos digitalizados (BRASIL, 2020). A Orientação Técnica nº 4 de 2016 do Conselho Nacional de Arquivos orienta que os documentos sejam "[...] produzidos diretamente em PDF/A ou em outros formatos e convertidos para PDF/A no momento do arquivamento" (CONARQ, 2016, p. 7).

A gestão de documentos é um “conjunto de medidas e rotinas que garante o efetivo controle de todos os documentos de qualquer idade, desde sua produção até sua destinação final” (BERNARDES, 1998). A relação do PDF/A com a gestão de documentos pode decorrer em todo o ciclo de vida do documento, como salienta o CONARQ (2016), “os documentos podem ser produzidos diretamente em PDF/A, ou produzidos em outros formatos e convertidos para PDF/A no momento do arquivamento”.

O padrão PDF/A mais básico, também conhecido por **padrão PDF/A-1**, foi posteriormente aprimorado para os padrões PDF/A-2, PDF/A-3 e PDF/A-4, como mostra a síntese do Quadro 1. O aprimoramento do padrão PDF/A para a versão 1.7, em 2011, denominado de **padrão PDF/A-2**, permitiu a inclusão de anexos no padrão PDF/A e é normalizado pela ISO 19005-2: 2011. Para a incorporação de anexos em qualquer formato foi criado o **padrão PDF/A-3**, definido pela ISO 19005-3: 2012. Esse padrão é amplamente adotado por assegurar que um determinado documento PDF/A, que é naturalmente lido por humanos, possa ter dados complementares em arquivos cujo formato possa ser também ser lido por máquina.

Quadro 1 - Padrões de PDF/A

Padrão de PDF/A	Documentação	Versão do PDF ao qual é baseado	Principais características
PDF/A-1	ISO 19005-1:2005	PDF 1.4 (Adobe Systems)	● Popularmente conhecido por PDF/A. ● Sendo o padrão mais usado atualmente, é gerado automaticamente pelos editores de

			texto populares e pelos programas de conversão.
PDF/A-2	ISO 19005-2:2011	PDF 1.7 (ISO 32000-1:2008)	• Compatível com o padrão PDF/A-1. • Aceita compactação JPEG 2000, camadas e transparência. • Aceita anexos padrão PDF/A.
PDF/A-3	ISO 19005-3:2012	PDF 1.7 (ISO 32000-1:2008)	• Compatível com o padrão PDF/A-2. • Permite quaisquer tipos de arquivos anexos.
PDF/A-4	ISO 19005-4:2019	PDF 2.0 (ISO 32000-2:2017)	• 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. • Aceita Javascript sem a execução pelo visualizador (por exemplo, para armazenar informações sobre a lógica de um formulário interativo).

Fonte: adaptado de PDFTron (2019a), LOC (2019a), LOC (2019b), Kimura e MAY (2019), LOC (2020a), LOC(2020b), LOC (2020c).

Como exemplo de uso do padrão PDF/A-3, o governo alemão emite faturas eletrônicas no padrão PDF/A legível por humanos e, para assegurar que seus dados fiquem bem representados em formato legível por máquina, usa o formato de arquivos XML inserindo-o no arquivo da fatura (PDFTRON, 2019a). Outra típica aplicação desse formato é na manufatura, onde há a incorporação da documentação 3D correspondente ao produto manufaturado na sua documentação (LOC, 2020b).

O desejo de compatibilidade entre os padrões PDF/A e PDF/E levou à criação do **padrão PDF/A-4** (LOC, 2020a), em 2020, para atender a demandas do campo da Engenharia. Diferentemente dos padrões anteriores, ele permite conteúdo não estático, como campos de formulários interativos implementados por meio da linguagem Javascript que, por questões de segurança, não podem ser executados pelo visualizador (LOC, 2020c).

Os padrões PDF/A-2, A-3 e A-4 estão em consonância com princípios da publicação ampliada, por permitirem a inserção de anexos. O padrão PDF/A, e suas variantes A-1, A-2 e A-3, é composto também por versões complementares denominadas de níveis de conformidade, representadas pelas letras 'a', 'b' e 'u', conforme mostra a síntese do Quadro 2. Observa-se que o padrão PDF/A-4 só permite os níveis 'f' e 'e'.

Quadro 2 - Níveis de conformidade do padrão PDF/A

Nível de Conformidade	Significado	Características	Padrões de PDF/A aplicáveis
a	Accesible (Acessível)	● Deve atender a todos os requisitos do padrão. ● Normalmente só é alcançado a partir de um documento nato-digital.	PDF/A-1 PDF/A-2 PDF/A-3
b	Basic (Básico)	● Nível mais baixo de conformidade. ● Requer apenas os padrões necessários para a reprodução do documento de forma confiável. ● Normalmente aplicado a documentos digitalizados.	PDF/A-1 PDF/A-2 PDF/A-3
u	Unicode[10]	● Nível intermediário de conformidade entre '**b**' e '**a**'. ● Todos os caracteres podem ser mapeados para o conjunto Unicode. ● Foi apresentado junto com o padrão PDF/A-2.	PDF/A-2 PDF/A-3
f	Files (Arquivos)	● Permite a incorporação de arquivos anexos em qualquer formato e atua como um sucessor do padrão PDF/A-3.	PDF/A-4
e	Engineering (Engenharia)	● É um sucessor do padrão PDF/E, direcionado a incorporação de anexos comuns da área da Engenharia em 3D e também arquivos Rich Media[11]. ● Aceita anexos de quaisquer formatos. ● Capaz de processar ações em Javascript se solicitado pelo usuário.	PDF/A-4

Fonte: adaptado de PDFTron (2019a), LOC (2019a), LOC (2019b), Kimura e May (2019), LOC (2020a), LOC(2020b), LOC (2020c).

2.3 PROBLEMAS, BENEFÍCIOS E RISCOS DO PDF/A E SEUS PADRÕES

Sendo o formato de arquivo PDF um formato praticamente onipresente no mundo dos documentos contemporâneos, a sua ampla adoção não implicou em uma diminuição aos riscos de preservação, porém, a maioria desses riscos foi eliminada no padrão PDF/A (WHEATLEY, 2019). Contudo, o padrão PDF/A é fundamentalmente um PDF e, portanto, herda os problemas derivados

[10] Unicode é um padrão para representar texto escrito em qualquer sistema linguístico e foi criado com o objetivo de transcender as limitações de codificações de caracteres tradicionais dos documentos digitais. Disponível em: https://home.unicode.org/.

[11] Rich Media representa um conjunto de arquivos com recursos avançados para vídeo, áudio ou outros elementos que incentivam os usuários a interagir e se envolver com o conteúdo. Disponível em: https://www.richmedia.com.

da sua complexidade e incerteza (KIMURA; MAY, 2019). Somando-se a isso, já foi constatado que muitos documentos digitais em formato PDF, identificados como sendo padrão PDF/A, advindos de ferramentas que afirmam ser capazes de criar PDF/A, não são totalmente compatíveis com esse padrão (LOC, 2020a). Dessa forma, há necessidade adicional de fazer a validação de um documento no padrão PDF/A com *softwares* específicos.

Kimura e May (2019) chamam atenção para um possível problema legal de direitos autorais com a criação de arquivos PDF/A quando as fontes são incorporadas, uma vez que elas são naturalmente incluídas no arquivo quando esse formato é adotado. Outro problema alertado pelos autores é que o padrão PDF/A é normalmente maior do que o seu correspondente PDF, uma vez que as fontes são incorporadas no arquivo. Além disso, o PDF/A pode crescer ainda mais quando adota os padrões PDF/A-2 ou PDF/A-3 devido a incorporação de anexos de tamanhos arbitrários.

Enquanto o padrão PDF/A-2 aceita somente anexos no formato de arquivo PDF/A, o padrão PDF/A-3 permite outros tipos de arquivos, gerando uma situação controversa em alguns casos e requerendo um manuseio cuidadoso e atenção especial (KIMURA; MAY, 2019), representando ainda um grande desafio para instituições de arquivamento, tanto para fluxos de trabalho quanto para gerenciamento e acesso de preservação a longo prazo (LOC, 2020b). É recomendável que se tenha uma política para controlar quais tipos de arquivos incorporados serão permitidos no padrão PDF/A-3 para minimizar seus riscos de preservação (PDFTRON, 2019b).

O relatório "The benefits and risks of the PDF/A-3 file format for archival Institution", elaborado por um grupo de trabalho da National Digital Stewardship Alliance (NDSA), abordou os principais riscos de uso do padrão PDF/A-3 em instituições arquivísticas (ARMS et al., 2014). Esse relatório reconhece a importância da possibilidade de se permitir a incorporação de tipos diversos de arquivos anexos para alguns cenários. Nesse caso, foi proposto o conceito de 'arquivamento híbrido', onde um documento fonte em um formato menos robusto à preservação é editado e depois incorporado em um arquivo PDF/A-3 e, assim, ter uma versão atualizada do documento em um formato robusto de preservação.

Contudo, para instituições de memória, tal uso de arquivos incorporados em documentos PDF/A-3 dependeria de protocolos muito específicos entre depositantes e repositórios de arquivos, onde "[...] os formatos aceitáveis como arquivos incorporados e com um fluxo de trabalho para garantir que o relacionamento entre o documento PDF e o arquivo incorporado seja

totalmente compreendido pela instituição arquivística" (ARMS et al., 2014, tradução nossa). Nesse sentido, os autores relatam uma preocupação no caso de um arquivo padrão PDF/A-3 possuir menor importância de preservação a longo prazo do que os seus documentos incorporados como anexos. Pois, nesse caso, os arquivos incorporados podem ter a sua preservação fragilizada uma vez que eles ficam sob a proteção de outro com menor importância no nível de preservação.

Quanto ao PDF/A-3 ter a possibilidade de conter conteúdo legível por humanos e dados anexados em formato legível por máquina, normalmente em formato de arquivo XML, pode trazer grande complexidade para os *softwares* leitores uma vez que esse tipo de articulação não faz parte do próprio padrão PDF (ARMS et al., 2014). Dessa forma, os autores recomendam que o padrão PDF/A-3 deve ser tratado com cautela e separadamente dos outros arquivos no padrão PDF/A, exigindo das instituições de memória um papel mais estratégico e ativo no processo de desenvolvimento de padrões.

Sobre a segurança, quanto a vírus ou programas indesejáveis, a especificação do padrão PDF/A-3, feita pela ISO 19005-3:2012, recomenda, mas não exige, que um *software* leitor só permita que a extração ou leitura de um anexo incorporado seja iniciada por uma ação explícita do usuário (ARMS et al., 2014). Assim, evita-se que usuários fiquem vulneráveis a problemas de segurança inerentes à abertura automática de arquivos desconhecidos, que não foram executados intencionalmente por eles.

Apesar dos riscos relatados, o uso do padrão PDF/A em conjunto com o seu recurso de inserção de anexos, permitido nas variantes A-2, A-3 e A-4, tem sido um grande aliado à publicação ampliada uma vez que esses anexos complementares ficam assegurados e agregados na publicação do documento PDF básico.

2.4 INCORPORAÇÃO DE METADADOS: DOCUMENTOS PDF AUTOREFERENCIADOS E AUTODOCUMENTADOS

No contexto da Arquivologia, metadados são “[...] dados estruturados que descrevem e permitem encontrar, gerenciar, compreender e/ou preservar documentos arquivísticos ao longo do tempo” (CONARQ, 2020), ou seja, eles encapsulam as informações que descrevem uma entidade portadora de informações (ZENG, 2020). Os metadados possuem vários níveis de granularidade e podem se referir a diferentes períodos do ciclo de vida dos dados, sendo incorporados no próprio documento a que se referem ou mantidos externos a ele (SAA, 2005a).

Prado Filho (2019) recomenda que os metadados também sejam gerenciados constituindo-se, assim, na gestão de metadados, considerada uma parte inseparável da gestão de documentos de arquivo. Prado Filho ainda afirma que os metadados nos "[...] processos de gestão de documentos de arquivo é um componente essencial para garantir a autenticidade, integridade, usabilidade e confiabilidade dos documentos de arquivo".

Apesar do formato de arquivo PDF ser considerado naturalmente ruim do ponto de vista semântico, ele tem evoluído muito nos últimos anos, conseguindo suportar a capacidade de incorporar metadados padronizados dentro do documento. Esses metadados podem ser usados por ferramentas de gerenciamento bibliográfico, mecanismos de pesquisa, ferramentas de mineração de texto etc. A vantagem mais evidente da incorporação de metadados em arquivos de formato PDF é que eles não correm o risco de serem separados do conteúdo do arquivo de forma involuntária ou por meio de uma má gestão documental.

Contudo, o grande desafio na incorporação de metadados em documentos PDF é a sua padronização semântica. Com essa preocupação, foi implementado na primeira versão do PDF oito metadados padrão, que existem até hoje (LEURS, 2020): (1) *Author* - quem criou o documento; (2) *CreationDate* - a data e hora em que o documento foi criado originalmente; (3) *Creator* - o aplicativo ou biblioteca de origem; (4) *Producer* - o produtor que criou o PDF; (5) *Subject* - sobre o que é o documento; (6) *Title* - o título do documento; (7) *Keywords* - as palavras-chave podem ser separadas por vírgula; (8) *ModDate* - a data e hora da última modificação.

De certa forma, pode-se considerar que um documento digital em formato PDF que tenha a maioria desses oito metadados básicos preenchidos é autodocumentado. Alguns desses metadados já são incorporados automaticamente no arquivo PDF, dependendo de qual *software* se utiliza, como a data de criação e o nome do autor. Outros metadados dessa lista podem ser editados conforme as funções oferecidas por cada editor. Atualmente existem possibilidades de inserir um conjunto muito maior de metadados em um arquivo PDF, inclusive de autoria do próprio usuário.

A partir da versão 1.4 do formato de arquivo PDF, foi criado um mecanismo capaz de incorporar metadados em arquivos PDF denominado de fluxo de metadados que se utiliza de uma estrutura chamada de XMP[12], normalizada pela ISO 16684-1:2019 (NISO, 2019). Esse fluxo é

[12] XMP (*extensible metadata platform*) é uma plataforma de metadados, desenvolvida pela Adobe, extensível que fornece um formato padrão para a criação, o processamento e o intercâmbio de metadados para uma grande

basicamente uma lista de pares com o nome do metadado e o seu valor. O padrão XMP foi projetado para ser extensível, pois além de permitir uso de metadados de vocabulários diversos, ele também permite adicionar um conjunto personalizado de metadados para satisfazer a necessidades bem específicas. Devido a essa flexibilidade e adaptabilidade o padrão XMP é amplamente usado no mundo por *softwares* e em diversos tipos de arquivos.

A LOC (2020a) recomenda a estrutura XMP como suporte à autodocumentação e afirma que essa estrutura é praticamente obrigatória para metadados básicos descritivos e de identificação. Devido a sua capacidade de trabalhar com vocabulários diversos o padrão XMP é bastante adotado e recomendado para muitas situações como, por exemplo, na incorporação de regras de *copyright* para documentos PDFs (CREATIVE COMMONS, 2015).

Existem diversas ferramentas que fazem a incorporação de metadados em arquivos PDF. A presente pesquisa selecionou uma delas para realizar o experimento. A metodologia usada na escolha é descrita na seção de procedimentos metodológicos e a forma de usá-la é explicitada na seção de resultados.

3 PROCEDIMENTOS METODOLÓGICOS

A pesquisa é exploratória descritiva e de levantamento bibliográfico. As etapas procedimentais foram: (i) pesquisa bibliográfica sobre a importância e viabilidade da incorporação de metadados e anexos em arquivos PDF; (ii) indicação de funções (ou critérios) para seleção das ferramentas candidatas; (iii) levantamento de ferramentas que atendessem aos critérios indicados; (iv) análise e escolha das ferramentas mais apropriadas para o contexto proposto na pesquisa; (v) criação de um fluxo procedimental, de fácil leitura, para auxiliar a execução das ações pelos estudantes com as ferramentas selecionadas; (vi) experimentação do fluxo procedimental sobre um conjunto delimitado de documentos.

Os documentos selecionados para o experimento pertencem ao acervo de trabalhos de conclusão de curso (TCC) da graduação em Arquivologia da Universidade Federal do Espírito Santo. O repositório de armazenamentos do acervo ainda é experimental e está em uma nuvem de acesso restrito que foi criada pela coordenação de TCCs. Cada registro de TCC do acervo permite a inserção de até 57 metadados, sendo organizados em oito grupos a saber: grupo 1 com três metadados de identificação do registro; grupo 2 com 28 metadados sobre o conteúdo; grupo

variedade de aplicativos. Disponível em: https://experienceleague.adobe.com/docs/experience-manager-cloud-service/content/assets/admin/xmp-metadata.html.

3 com dois metadados sobre a autoria; grupo 4 com dois metadados sobre a orientação; grupo 5 com oito metadados sobre a avaliação e defesa; grupo 6 com três metadados sobre a forma do documento; grupo 7 com oito metadados sobre o suporte e digitalização; e grupo 8 com quatro metadados sobre os direitos e disseminação. Ainda há duas tabelas de dados para apoio: uma específica para o vocabulário controlado de autoridades e outra para o vocabulário controlado das instituições.

Para a seleção inicial das ferramentas, foi feita uma busca na Web, usando o termo base "PDF" intercalando-o com os termos específicos: "editor", "leitor/reader", "metadado/metadata", "anexo/attachment", "validação/validation", "validador/validator", "conformidade/conformity", "metadado/metadata" e "incorporar/embed".

Para a classificação das ferramentas candidatas à análise, foram elaboradas 12 funções sobre aspectos relevantes ao contexto da pesquisa e considerando elementos importantes apresentados na revisão bibliográfica da seção 2. São eles: (i) dá acesso gratuito às funções analisadas; (ii) exibe documentos PDF/A-1; (iii) exibe documentos PDF/A-2, A-3 e A-4, e seus anexos; (iv) insere anexos em formato de arquivo PDF/A para satisfazer o padrão PDF/A-2; (v) insere anexos em qualquer formato para satisfazer os padrões PDF/A-3 e A-4; (vi) exibe e valida o nível de conformidade de um PDF; (vii) define o padrão e o nível de conformidade do PDF/A para A-1, A-2 e A-3; (viii) define o padrão e o nível de conformidade do PDF/A para PDF/A-4; (ix) exibe os metadados básicos de um PDF; (x) exibe quaisquer metadados padrão XMP de um PDF; (xi) incorpora e edita metadados básicos em um PDF; (xii) incorpora e edita quaisquer metadados padrão XMP em um PDF.

Os testes, verificações e execuções, quanto ao atendimento das funções, foram realizadas utilizando-se o sistema operacional Windows 10, por ser mais acessível aos alunos.

Para representação do fluxo procedimental, usou-se a BPMN (*business process model and notation*) com suporte da ferramenta de desenho Draw.io[13]. A BPMN é uma linguagem notacional para representar de forma gráfica e com muita expressividade os processos que ocorrem em praticamente todo tipo de organização (CHINOSI; TROMBETTA, 2012).

[13] Draw.io é uma ferramenta para desenho de diagramas com suporte a vários conjuntos notacionais conforme a área de conhecimento. Ela é multiplataforma, gratuita e executada diretamente no navegador Web. Disponível em: https://drawio-app.com/.

4 RESULTADOS

Essa seção apresenta as ferramentas candidatas resultantes do levantamento e as três ferramentas escolhidas conforme os critérios estabelecidos. Apresenta também um fluxo procedimental, como sugestão para uso das ferramentas escolhidas no processo de depósito do TCC no acervo, demonstrando o seu uso com um exemplo. Finalmente, discute a pertinência dos resultados com os princípios FAIR e a publicação ampliada.

4.1 FERRAMENTAS ESCOLHIDAS PARA INCORPORAÇÃO DE ANEXOS E METADADOS E VALIDAÇÃO DO PADRÃO PDF/A

O Quadro 3 apresenta as 16 ferramentas que atenderam a pelo menos quatro dos critérios, conforme mostra a última linha do Quadro 4. As colunas do Quadro 3 são: (i) número da ferramenta, usado na referência no Quadro 4; (ii) nome da ferramenta; (iii) o endereço Web onde se encontra disponível; (iv) o ano/número da versão analisada, onde 'ON' indica ser uma versão online executada em um navegador Web, e 'OFF' indica que a versão precisa ser instalada no computador.

Quadro 3 - Ferramentas candidatas para a análise.

Id.	**Nome da ferramenta** (em ordem alfabética)	**Disponível em:**	**Versões analisadas, online (ON)/ offline (OFF)**
01	Adobe Acrobat Professional	https://www.adobe.com/br/acrobat/acrobat-pro.html	XI OFF
02	Adobe Acrobat Reader	https://www.adobe.com/br/acrobat/pdf-reader.html	2021 OFF
03	ASPOSE	https://products.aspose.app/pdf	2021 ON
04	AvePDF	https://avepdf.com/pt	2021 ON
05	Exiftool	https://exiftool.org	2021 OFF
06	Foxit Editor	https://www.foxit.com/pdf-editor	2021 OFF
07	Foxit Reader	https://www.foxitsoftware.com/pdf-reader	2021 OFF
08	GROUPDOCS	https://products.groupdocs.app	2021 ON
09	I Love PDF	https://www.ilovepdf.com	2021 ON OFF
10	PDF Candy	https://pdfcandy.com/pt	2021 ON OFF
11	Pdf Metadata Editor	https://github.com/zaro/pdf-metadata-editor	2021 OFF
12	Pdf Wondershare	https://pdf.wondershare.com.br	2021 OFF
13	PDFCreator Free	https://www.pdfforge.org/pdfcreator	4.4.1 OFF
14	pdfmark	https://github.com/Crossref/pdfmark	2018 OFF
15	PDFTron Demo	https://www.pdftron.com/pdf-sdk	2021 ON

16	Sedja	https://www.sejda.com	2021 ON OFF

Fonte: autoria própria.

O Quadro 4 contém uma síntese do resultado da análise comparativa das 16 ferramentas apresentadas no Quadro 3, mostrando o atendimento (S) ou não (N) de cada uma das 12 funções.

Quadro 4 - Resultado da análise realizada sobre as ferramentas do Quadro 3

Função analisada	**01**	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
i) Dá acesso gratuito às funções analisadas	N	S	S	S	S	N	S	S	N	S	S	N	S	S	S	S
ii) Exibe documentos PDF/A-1	S	S	S	N	N	S	S	N	S	S	N	S	N	N	S	S
iii) Exibe documentos PDF/A-2, A-3 e A-4, e seus anexos	S	S	N	N	N	S	S	N	S	N	N	S	N	N	N	N
iv) Insere anexos em formato de arquivo PDF/A para satisfazer o padrão PDF/A-2	S	S	N	N	N	S	S	N	N	N	N	S	N	N	N	N
v) Insere anexos em qualquer formato para satisfazer os padrões PDF/A3 e A4	S	S	N	N	N	S	S	N	N	N	N	S	N	N	N	N
vi) Exibe e valida o nível de conformidade de um PDF	N	N	N	S	N	N	N	N	N	N	N	N	N	N	S	N
vii) Define o padrão e o nível de conformidade do PDF/A para A-1, A-2 e A-3	N	N	N	S	N	N	N	N	S	N	N	N	S	N	S	N
viii) Define o padrão e o nível de conformidade do PDF/A para PDF/A-4	N	N	N	S	N	N	N	N	N	N	N	N	N	N	N	N
ix) Exibe os metadados básicos de um PDF	S	S	S	S	S	S	S	S	S	S	S	S	S	S	N	S
x) Exibe quaisquer metadados padrão XMP de um PDF	S	N	N	N	S	S	N	S	N	N	S	N	N	S	N	N
xi) Incorpora e edita metadados básicos em um PDF	S	N	S	S	S	S	S	S	N	S	S	S	S	S	N	S
xii) Incorpora e edita quaisquer metadados padrão XMP em um PDF	S	N	N	N	S	S	N	N	N	N	S	N	N	S	N	N
Quantidade de características assinaladas com 'SIM'	**8**	**6**	**4**	**6**	**5**	**8**	**7**	**4**	**4**	**4**	**5**	**6**	**4**	**5**	**4**	**4**

Fonte: autoria própria.

Além da quantidade de funções atendidas, mostrada na última linha do Quadro 4, foram considerados mais fortemente, os critérios de gratuidade, completude de disponibilização de recursos para trabalhar os padrões A-1, A-2 e A-3, usabilidade e popularidade.

As duas ferramentas com melhor desempenho, no quesito quantidade de funções atendidas, o Adobe Acrobat Professional e o Foxit Editor, ambas com oito funções, não oferecem acesso gratuito a todas as funções analisadas. Foram então descartadas, pois o aspecto da gratuidade é necessário no contexto da comunidade de usuários da pesquisa.

O Quadro 5 mostra as três ferramentas escolhidas, após os testes e análise realizada. Para a incorporação de anexos, foi escolhida a ferramenta Foxit Reader, que é um *software* gratuito, popular e com boa usabilidade e obteve sete funções atendidas. Quanto a transformação para o padrão PDF/A e suas variantes, bem como para a sua validação, foi escolhida a ferramenta AvePDF, gratuita que funciona direto no navegador Web e com boa usabilidade, também mostrada no Quadro 5.

Quadro 5 - Ferramentas escolhidas.

Propósito	Ferramenta	Link de acesso
Incorporação de anexos	Foxit Reader	https://www.foxitsoftware.com/pdf-reader
Validação/transformação para o padrão PDF/A	AvePDF	https://avepdf.com/pt
Incorporação de metadados	Exiftool	https://exiftool.org

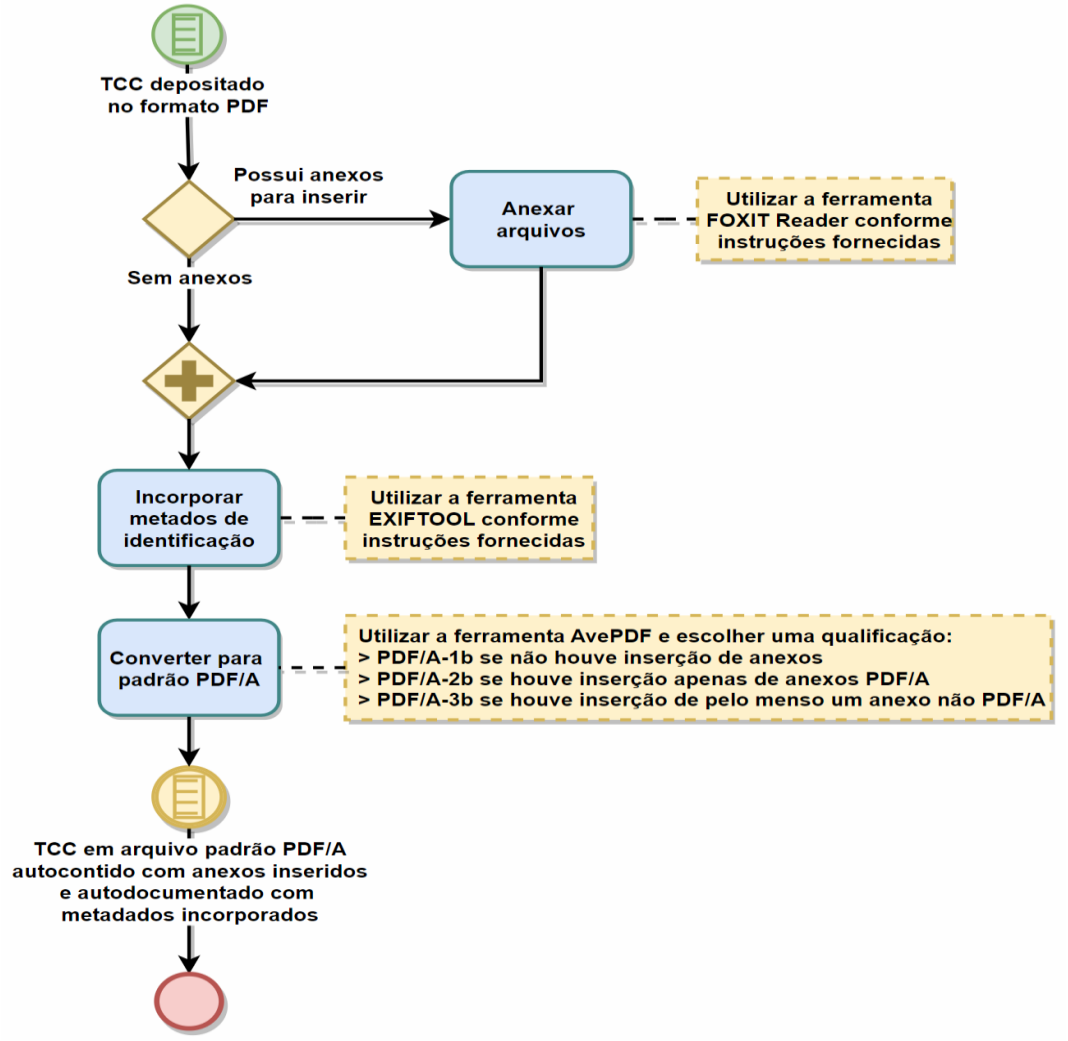
Fonte: autoria própria.

Para a incorporação de metadados em arquivo PDF existem 13 opções, dentre as ferramentas selecionadas no Quadro 3, quando se trata dos metadados básicos. Porém, poucas opções de ferramentas que trabalham com quaisquer metadados no formato XMP, e menos ainda quando considerado o critério de gratuidade do uso. Nesse caso, foram apenas três ferramentas: Exiftool, Pdf Metadata Editor, e pdfmaker. Apesar da ferramenta Exiftool não possuir boa usabilidade, pois trabalha com linha de comando, ela tem vantagem sobre as outras duas por permitir a serialização no processo de inserção dos metadados. Assim, a ferramenta Exiftool foi escolhida para incorporação de metadados, conforme mostra o Quadro 5.

4.2 FLUXO PROCEDIMENTAL

O fluxo procedimental, em linguagem BPMN, para realizar as ações necessárias é mostrado na Figura 1. O fluxo mostra a transformação do documento principal do TCC em formato de arquivo PDF para um documento autocontido e autodocumentado, isto é, no padrão PDF/A qualificado e com anexos complementares à pesquisa, e ainda com metadados semânticos de identificação incorporados. A característica autorreferenciado não foi citada no fluxo pois a criação dos identificadores ISBN, ISSN e DOI não foi feita na presente pesquisa por exigir a alocação de recursos financeiros.

Figura 1 - Fluxo procedimental para depósito do TCC no acervo.



Fonte: autoria própria.

O fluxo é atualmente utilizado por um estagiário de Arquivologia na Coordenação de TCCs do curso, com auxílio de instruções específicas de operação de cada um dos três *softwares*

escolhidos e apresentados no Quadro 5 da seção 4.1. Os próximos parágrafos demonstram um uso do fluxo procedimental da Figura 1.

A demonstração supõe um TCC cujo conteúdo está representado em um arquivo PDF denominado de "**tcc_fulano_de_tal.pdf**" e que a sua respectiva pesquisa acadêmica gerou um conjunto de dados armazenado no arquivo "**dados.csv**". Apesar dos 57 metadados possíveis associados a um TCC do acervo, são demonstrados apenas três, indicados no Quadro 6.

Seguindo o fluxo procedimental da Figura 1, o arquivo "**dados.csv**" é inserido no arquivo "**tcc_fulano_de_tal.pdf**" por meio da ferramenta Foxit Reader. Em seguida, os metadados do Quadro 6 são incorporados no arquivo "**tcc_fulano_de_tal.pdf**" com auxílio da ferramenta Exiftool cuja execução está demonstrada no Quadro 7.

Quadro 6 - Metadados do TCC da demonstração.

Metadado	Valor do metadado
titulo	Preservação digital
autor	José da Silva
semestreLetivo	2022-1

Fonte: autoria própria.

O último passo do fluxo é a conversão do PDF para o padrão PDF/A por meio da ferramenta AvePDF. Nesse caso, como foi inserido um arquivo anexo no formato CSV, a qualificação escolhida no AvePDF é a PDF/A-3b.

Quadro 7 - Execução da ferramenta Exiftool.

Arquivo: metadados.cfg

```
%Image::ExifTool::UserDefined = ('Image::ExifTool::XMP::pdfx' => {
titulo => {},
autor => {},
semesteLetivo => {},
},);
```

Arquivo: valores.txt

```
-titulo=Preservação digital
-autor=José da Silva
-semestreLetivo=2022-1
```

Comando de execução no prompt:

```
exiftool -config metadados.cfg -charset utf8 -charset iptc=utf8 -
codedcharacterset=utf8 -@ valores.txt "tcc_jose_da_silva.pdf"
```

Fonte: autoria própria.

4.3 PERTINÊNCIA DOS RESULTADOS COM OS PRINCÍPIOS FAIR E A PUBLICAÇÃO AMPLIADA

Os princípios FAIR prezam por quatro capacidades relativas aos dados: (i) *findable* - capacidade de serem encontrados, (ii) *accessible* - capacidade de serem acessíveis; (iii) *interoperable* - capacidade de serem interoperáveis e (iv) *reusable* - capacidade de serem reutilizáveis. Essas quatro capacidades têm forte relação com as propriedades dos documentos de serem autocontidos e autodocumentados, trabalhadas no experimento,.

O conteúdo de um documento autocontido PDF/A possui boa capacidade de acesso e reuso devido a sua portabilidade em vários sistemas e softwares e pela independência de elementos externos. Os dados de pesquisa anexados em um documento PDF/A-3 atendem aos princípios *accessible* e *reusable* pois eles ficam disponíveis no próprio documento. Os metadados incorporados em documentos padrão PDF/A autodocumentados atendem ao princípio *findable* pois ficam mais fáceis de serem encontrados em comparação com a situação de estarem apenas dentro do texto.

A adoção do padrão XMP vai ao encontro da necessidade do uso da interoperabilidade para garantir "[...] o uso e a encontrabilidade dos metadados estruturados dos objetos informacionais" (LUZ, 2018). Com o uso do padrão XMP e também de vocabulários padronizados e conhecidos é possível aumentar o nível de interoperabilidade dos metadados incorporados de um documento autodocumentado. Especificamente quanto a tarefa do uso de vocabulários padronizados para associar aos metadados existentes, ainda não foi desenvolvida na presente pesquisa.

É importante destacar que o arquivista, juntamente com outros profissionais, deve ser atuante na gestão de metadados, conforme orienta a norma técnica NBR ISO 2381-1:2019, onde é recomendado que "[...] os arquivistas, decidam sobre como armazenar os metadados [...]" e como será "[...] a ligação permanente entre os metadados e os objetos aos quais eles se relacionam ou pertencem [...]" podendo ter o armazenamento de ambos juntos (ABNT NBR ISO, 2019). A parte do experimento que incorpora metadados em um arquivo PDF/A mostrou que essa atuação desejada do arquivista é factível.

Quanto a publicação ampliada é possível inferir que o padrão PDF/A-3, por permitir a inserção de arquivos de diversos formatos, cumpre requisitos importantes dessa modalidade de publicação. No caso do experimento demonstrado, o arquivo "`dados.csv`" possui informações referentes à pesquisa relatada no conteúdo do documento PDF e, após a sua inserção, os dois

itens informacionais referentes à pesquisa, o texto e os dados, ficam vinculados no próprio arquivo. Assim, quando esse for publicado, o acesso aos seus dados estará facilitado, uma vez que leitores de PDF normalmente exibem arquivos anexos. Outro aspecto que faz suporte à publicação ampliada é a ação de incorporação de metadados semânticos de identificação no documento PDF/A.

5 CONSIDERAÇÕES FINAIS

Documentos digitais em formato PDF e caracterizados como autocontidos, padrão PDF/A-1, são importantes na preservação por não precisarem de elementos externos para a sua exibição. Quando possuem arquivos anexos são importantes como suporte à publicação ampliada devido a vinculação do seu conteúdo com seus dados complementares. Se existem metadados de identificação única incorporados, como o ISBN, ISSN ou DOI, são caracterizados como autorreferenciados, ou ainda, se os metadados incorporados servem para identificação complementar, são caracterizados como autodocumentados. Em ambos casos de incorporação de metadados há suporte à publicação ampliada, pois ocorre a agregação do conteúdo do PDF com metadados que trazem informações sobre o documento. Além disso, nesses casos de incorporação de metadados, há o atendimento aos princípios FAIR.

Houve uma investigação sobre os elementos conceituais e as ferramentas computacionais apropriadas para criação de documentos PDF autocontidos, autorreferenciados e autodocumentados de forma a atender aos princípios FAIR e em um contexto de publicação ampliada. Após um levantamento na Web e o estabelecimento de critérios, foram escolhidas as ferramentas Foxit Reader e Exiftool para incorporar anexos e metadados, respectivamente, em arquivos de formato PDF, e a ferramenta AvePDF, para validação de arquivos no padrão PDF/A. Em função dessas escolhas, foi apresentado, na subseção 4.2, um fluxo procedimental para adequar os documentos digitais em formato PDF, oriundos de TCCs, para se tornarem autocontidos e autodocumentados, usando ferramentas computacionais escolhidas. Dessa forma, os dois objetivos elencados na presente pesquisa foram atingidos.

A característica de documento autorreferenciado, apesar de ter sido conceitualmente trabalhada na pesquisa, não foi considerada factível para o experimento por envolver recursos financeiros para a criação dos identificadores ISBN, ISSN e DOI. Contudo, o processo de incorporação desses identificadores é aquele já demonstrado para a incorporação dos metadados por meio do padrão XMP, uma vez que eles são também considerados metadados.

Os padrões PDF/A-3 e PDF/A-4 expressam de forma plena a representação dos novos paradigmas elucidados por Cook (2012), abrindo opções arquivísticas para se trabalhar com diversos documentos e com a incorporação facilitada da organização de acervos de documentos relacionados. Quanto ao padrão mais novo PDF/A-4, a Library of Congress ainda não expressou preferência pelo seu uso, pois aguarda o feedback da experiência de uso da comunidade (LOC, 2020c).

Quanto ao cenário usado nos experimentos, a Portaria MEC nº 1.224 de 2013 instituiu normas sobre a manutenção e guarda do Acervo Acadêmico das Instituições de Educação Superior (IES) pertencentes ao sistema federal de ensino (BRASIL, 2013), aprovada também pelo Arquivo Nacional, sugerindo a eliminação como destinação final aos TCCs. Sugere também a inexistência da fase intermediária e a devolução ao aluno após o registro das notas, na fase corrente. Contudo, a portaria não proíbe que os TCCs sejam guardados na instituição, como ocorrido em diversas IES, inclusive na coordenação do curso onde os experimentos foram realizados. Contudo, acredita-se que os resultados da presente pesquisa possam ser mais significativos em contextos informacionais com elementos de maior temporalidade de guarda ou caracterizados como preservação permanente como ocorre, por exemplo, com dissertações de mestrado e teses de doutorado. Nesses contextos, o suporte à publicação ampliada seria mais fortemente caracterizado.

Como continuidade da pesquisa, sugere-se a melhoria da interoperabilidade associando cada metadado do TCC a um metadado correspondente pertencente a vocabulários padronizados que possuam ampla abrangência de uso nos repositórios de trabalhos acadêmicos.

REFERÊNCIAS

ABNT NBR ISO. ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR ISO 19005-1:2009** - Gerenciamento de documentos — formato eletrônico de arquivo de documento para preservação de longo prazo, parte 1: uso do PDF 1.4 (PDF/A-1). ISBN 978-85-07-01246-7. 23 fev. 2009.

ABNT NBR ISO. ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR ISO 23081-1:2019** - Informação e documentação - Processos de gestão de documentos de arquivo - Metadados para documentos de arquivo. Parte 1: Princípios. Rio de Janeiro. 19 set. 2019.

ADOBE ACROBAT. **O que é um PDF?**. 2022. Disponível em: https://www.adobe.com/br/acrobat/about-adobe-pdf.html. Acesso em: 22 ago. 2022.

ARQUIVO NACIONAL. Política de preservação digital. **Arquivo Nacional**, v. 2. 2016. Disponível em: http://www.siga.arquivonacional.gov.br/images/an_digital/and_politica_preservacao_digital_v2.pdf. Acesso em: 22 ago. 2022.

BRASIL. Portaria MEC nº 1.224, de 18 de dezembro de 2013. Institui normas sobre a manutenção e guarda do Acervo Acadêmico das Instituições de Educação Superior (IES) pertencentes ao sistema federal de ensino. **Diário Oficial da União**, Brasília, DF, 18 dez. 2013.

BRASIL. Decreto nº 10.278, de 18 de março de 2020. Regulamenta o disposto no inciso X do caput do art. 3º da Lei nº 13.874, de 20 de setembro de 2019, e no art. 2º-A da Lei nº 12.682, de 9 de julho de 2012... **Diário Oficial da União**. 2020.

CÂMARA TÉCNICA DE DOCUMENTOS ELETRÔNICOS. **Glossário**. Disponível em https://www.gov.br/conarq/pt-br/assuntos/camaras-tecnicas-setoriais-inativas/camara-tecnica-de-documentos-eletronicos-ctde/glosctde_2020_08_07.pdf. Acesso em: 22 ago. 2022.

CHINOSI, Michele; TROMBETTA, Alberto. BPMN: An introduction to the standard. **Computer Standards & Interfaces**, v. 34, n. 1, p. 124–134, 2012. DOI: 10.1016/j.csi.2011.06.002. Disponível em: https://www.sciencedirect.com/science/article/pii/S0920548911000766. Acesso em: 22 ago. 2022.

CONARQ. **Orientação técnica nº 4**: recomendações de uso do PDF/A para documentos arquivísticos. 2016. Disponível em: http://conarq.gov.br/images/ctde/Orientacoes/Orientacao_tecnica_4.pdf. Acesso em: 22 ago. 2022.

CONARQ. **Glossário**: documentos arquivísticos digitais (8ª versão). Rio de Janeiro: Conselho Nacional de Arquivos - CONARQ/Câmara Técnica de Documentos Eletrônicos - CTDE, 2020. Disponível em: https://www.gov.br/conarq/pt-br/assuntos/camaras-tecnicas-setoriais-inativas/camara-tecnica-de-documentos-eletronicos-ctde/glosctde_2020_08_07.pdf. Acesso em: 22 ago. 2022.

COOK, Terry. Arquivologia e pós-modernismo: novas formulações para velhos conceitos. **Informação Arquivística**, v. 1, n. 1, p. 26, 2012. Disponível em: https://www.brapci.inf.br/_repositorio/2015/12/pdf_deb3461ca4_0000018241.pdf. Acesso em: 22 ago. 2022.

CREATIVE COMMONS. **Embedded metadata with XMP**. 2015. Disponível em: https://wiki.creativecommons.org/images/f/ff/Creativecommons-embedded-metadata-with-xmp_eng.pdf.

DURANTI, Luciana. Registros documentais contemporâneos como provas de ação. **Revista Estudos Históricos**,v. 7, n. 13, p. 49–64, 1994. Disponível em: http://bibliotecadigital.fgv.br/ojs/index.php/reh/article/view/1976. Acesso em: 22 ago. 2022.

KIMURA, Akiko; MAY, Peter. **PDF format preservation assessment, part 2**: PDF/A profile. British Library, 2019. Disponível em: https://wiki.dpconline.org/images/2/22/PDFA_Assessment_v1.0.pdf. Acesso em: 22 ago. 2022.

LOC. PDF/A-1, PDF for long-term preservation, use of PDF 1.4. **Sustainability of Digital Formats**: Planning for Library of Congress Collections. 2019a. Disponível em: https://www.loc.gov/preservation/digital/formats/fdd/fdd000125. Acesso em: 22 ago. 2022.

LOC. PDF/A-2, PDF for long-term preservation, use of ISO 32000-1 (PDF 1.7). **Sustainability of Digital Formats**: Planning for Library of Congress Collections. 2019b. Disponível em: https://www.loc.gov/preservation/digital/formats/fdd/fdd000319. Acesso em: 22 ago. 2022.

LOC. PDF/A Family, PDF for long-term Preservation. **Sustainability of Digital Formats**: Planning for Library of Congress Collections. 2020a. Disponível em:

https://www.loc.gov/preservation/digital/formats/fdd/fdd000318. Acesso em: 22 ago. 2022.

LOC. PDF/A-3, PDF for long-term preservation, use of ISO 32000-1, with embedded files. **Sustainability of Digital Formats**: Planning for Library of Congress Collections. 2020b. Disponível em: https://www.loc.gov/preservation/digital/formats/fdd/fdd000360. Acesso em: 22 ago. 2022.

LOC. PDF/A-4, PDF for long-term preservation, use of ISO 32000-2. **Sustainability of Digital Formats**: Planning for Library of Congress Collections. 2020c. Disponível em: https://www.loc.gov/preservation/digital/formats/fdd/fdd000532. Acesso em: 22 ago. 2022.

LOC. PDF 2.0, ISO 32000-2 (2017, 2020). **Sustainability of Digital Formats**: Planning for Library of Congress Collections. 2020d. Disponível em: https://www.loc.gov/preservation/digital/formats/fdd/fdd000474. Acesso em: 22 ago. 2022.

NISO. **ISO 16684-1:2019** graphic technology, extensible metadata platform (XMP) specification, part 1: data model, serialization and core properties. 2019. Disponível em: https://www.iso.org/standard/75163.html. Acesso em: 22 ago. 2022.

PDFTTRON. **What are the different versions of PDF/A?**. 2019a. Disponível em: https://pdftron.com/blog/pdfa-format/what-are-the-different-types-of-pdfa. Acesso em: 22 ago. 2022.

PDFTRON. **How to pick the right version of PDF/A**. 2019b. Disponível em: https://pdftron.com/blog/pdfa-format/how-to-pick-right-version-of-pdfa. Acesso em: 22 ago. 2022.

RODRIGUES, Fernando de Assis; SANT'ANA, Ricardo César Gonçalves. Publicação Ampliada no Contexto de Teses e Dissertações. **Informação & Tecnologia**, Marília/João Pessoa, v. 3, n. 1, p. 4–26, 2016. Disponível em: https://periodicos.ufpb.br/ojs/index.php/itec/article/view/38248. Acesso em: 22 ago. 2022.

RONDINELLI, R. C. **O documento arquivístico ante a realidade digital:** uma revisão conceitual necessária. Rio de Janeiro: FGV, 2013.

SAA. Portable Document Format (PDF). In: **Archives Teminology**, Society of American Archivists. 2005a. Disponível em: https://dictionary.archivists.org/entry/portable-document-format.html. Acesso em: 22 ago. 2022.

SAA. Metadata. In: **Archives Teminology**, Society of American Archivists (SAA). 2005b. Disponível em: https://dictionary.archivists.org/entry/portable-document-format.html. Acesso em: 22 ago. 2022.

SALES, Luana Farias; DE SOUZA, Rosali Fernandez; SAYÃO, Luís Fernando. Publicação ampliada: um novo modelo de publicação científica voltada para os desafios de uma ciência orientada por dados. *In:* ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 15., 2014. **Anais eletrônicos**... Em. Belo Horizonte, 27 a 31 de outubro de 2014. Disponível em: https://ridi.ibict.br/handle/123456789/824. Acesso em: 22 ago. 2022.

SAYÃO, Luís Fernando; SALES, Luana Farias. Curadoria digital e dados de pesquisa. **AtoZ**: novas práticas em informação e conhecimento, v. 5, n. 2, p. 67–71, 2016. DOI: 10.5380/atoz.v5i2.49708. Disponível em: https://revistas.ufpr.br/atoz/article/view/49708. Acesso em: 22 ago. 2022.

VASILESCU, Ramona. PDF/A standard for long term archiving. **Computer Science Series**, v. 7, n. 1, p. 8, 2009. Disponível em: https://arxiv.org/ftp/arxiv/papers/0906/0906.0867.pdf. Acesso em: 22 ago. 2022.

WHEATLEY, Paul; MAY, Peter; PENNOCK, Maureen; KIMURA, Akiko; WHIBLEY, Simon; RUSSO, David. **PDF**

format preservation assessment, part 1: PDF. British Library, 2019. Disponível em: https://wiki.dpconline.org/images/f/ff/PDF_Assessment_v1.5.pdf. Acesso em: 22 ago. 2022.

ZENG, Marcia L. **Metadata basics**. 2020. Disponível em: http://metadataetc.org/metadatabasics/. Acesso em: 22 ago. 2022.

NOTAS DE AUTORIA

Henrique Cristovão
Doutor em Ciência da Informação na Universidade de Brasília (UnB) com estágio de pesquisa (Doutorado Sanduíche) no Institute for Human & Machine Cognition (IHMC/EUA). Mestre em Informática na Universidade Federal do Espírito Santo (UFES). Bacharel em Matemática Aplicada e Computacional na UFES. Professor Adjunto na UFES, Departamento de Arquivologia. Líder do grupo de pesquisa Organização e Recuperação de Conhecimento em Rede (NetKOR), registrado no CNPq. Participa dos grupos de pesquisa: Observatório da Informação Arquivística Digital, Tecnologias da Informação e Comunicação Aplicadas à Saúde, e Inteligência Cooperativa em Redes Sociais Complexas. Editor do Brazilian Journal of Production Engineering (BJPE) na área Informação e Conhecimento. Membro da Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação (ANCIB). Linhas de pesquisa atuais: Networked Knowledge Organization Systems, Knowledge Discovery (NKOS), Web Semântica e Linked Open Data, Modelagem Conceitual, Análise de Redes Complexas, e Organização e Gestão do Conhecimento.

Link Currículo Lattes - http://lattes.cnpq.br/5035919384923489

Willian Alves Batista
Bacharel em Arquivologia na Universidade Federal do Espírito Santo (UFES). Projetos desenvolvidos, apoio na organização de arquivos institucionais na Secretaria de Avaliação Institucional da UFES e Monitoria especial e temporária para o Centro de Ciências Jurídicas e Econômicas da UFES.

Link Currículo Lattes - http://lattes.cnpq.br/9673198241662286

Bruna Morêto Sibaldo Rocha
Bacharelanda em Arquivologia na Universidade Federal do Espírito Santo (UFES). Iniciação científica na área de metadados de arquivos por um ano. Monitoria na matéria de Preservação e Conservação de Documentos I.

Link Currículo Lattes - http://lattes.cnpq.br/2366566452976746