INTEGRIDADE DE DOCUMENTOS DIGITAIS ENTRE SIGAD E RDC-ARQ

Carlos Eduardo Carvalho Amand (Faculdade FAVENI/ Universidade do Estado do Rio de Janeiro), Brenda Couto de Brito Rocco (Universidade do Estado do Rio de Janeiro)

1 INTRODUÇÃO

O advento do documento digital, cada vez mais, vem tomando espaço no mundo e, por consequência, no Estado Brasileiro, intensificando-se, atualmente, com a "transformação digital" fomentada pelo governo eleito em 2014. A produção de documentos digitais, como meio preferencial de construção de processos administrativos, teve como grande marco a publicação do Decreto nº 8.539, de outubro de 2015, pela então presidente Dilma Rousseff, pois, com ele, veio o reconhecimento dos documentos natos digitais como originais para todos os efeitos, igualando-os aos documentos convencionais em suporte de papel.

O decreto também impulsionou o uso do meio digital para a produção do processo administrativo, impondo às entidades funcionais do Estado um prazo para implantação do processo digital em dois anos ou, no máximo, em três anos no caso de órgãos ou entidades que já utilizassem esse processo eletrônico.

Ainda em outubro de 2015, os ministérios da justiça e do planejamento, orçamento e gestão emitem a Portaria Interministerial nº 1.677, que define os

procedimentos de protocolo e de observância obrigatória aos órgãos e entidades da Administração Pública Federal, bem como os procedimentos para tratar documentos avulsos e processos no suporte convencional e no suporte digital são detalhados. Há que se destacar que, nesta Portaria, institui-se a obrigatoriedade para os órgãos e entidades da Administração Pública Federal observarem, em seus sistemas de protocolo, a aderência aos requisitos descritos no Modelo de Requisitos para Sistemas Informatizados de Gestão Arquivística de Documentos (e-ARQ Brasil).

O e-ARQ Brasil é um modelo de requisitos para a construção de Sistemas Informatizados de Gestão Arquivísticas de Documentos – SIGAD, emitido pelo Conselho Nacional de Arquivos, doravante Conarq, com sua edição final em 2011[157]. Este modelo reúne as melhores práticas de gestão arquivista tendo como base normas e padrões internacionais, como DOD 5010.2-ST[158] e PREMIS[159].

O Conarq publicou, em 04 de setembro de 2015, a Resolução nº 43, que estabelece diretrizes para a implementação de repositórios digitais confiáveis. Essas diretrizes têm por objetivo manter a autenticidade, a confidencialidade, a disponibilidade, o acesso e a preservação dos documentos arquivísticos digitais.

O uso de documentos digitais se intensifica, ainda mais, com a emissão do Decreto Federal nº 10.278, de 18 de março de 2020, que estabelece a técnica e os requisitos para a digitalização de documentos públicos e privados de forma que os digitalizados possuam o mesmo efeito legal que os documentos originais. O decreto também permite o descarte do documento original que foi digitalizado, exceto quando eles forem de valor permanente e, segundo o decreto, possuírem "valor histórico".

Com a sociedade passando a produzir, em massa, documentos natos digitais e procedendo com a digitalização de documentos convencionais, faz-se necessário garantir suas integridades quando houver a troca de custódia legal, que pode ocorrer nos processos de transferência/recolhimento ou no envio de correspondências entre entidades. Assim, este artigo propõe uma solução tecnológica para a preservação da integridade de documentos digitais durante a execução desses processos. Consideramos, para fins desse estudo, integridade como o "estado dos documentos que se encontram completos e não sofreram

[157] CONARQ, Resolução nº 25, de 27 de abril de 2007. Recentemente revogada pela Resolução nº 50, de 06 de maio de 2022, cujo anexo é a versão 2 do e-Arq Brasil.
[158] Digital criteria standard for eletronic record managment software applications: DOD 5015.2-STD, 2002.
[159] PREMIS Data Dictionary for Preservation Metadata – Version 2.

nenhum tipo de corrupção ou alteração não autorizada nem documentada", conforme Conarq (2015, p. 7).

2 DESENVOLVIMENTO

Para um melhor entendimento da solução proposta, é necessária uma revisão dos conceitos que serão estudados. Vamos explorar o ciclo de vida de documentos e sua aplicação desde o SIGAD até o RDC-Arq, objetivando entender o processo de transferência/recolhimento, bem como revisar o padrão de empacotamento BagIt para a formação de pacotes de transferência e, por fim, explicar a técnica de função criptográfica *hash*.

3 CICLO VITAL DE DOCUMENTOS ARQUIVÍSTICOS

O Arquivo Nacional (BRASIL, 2005, p. 160) define ciclo vital de documentos como as "sucessivas fases por que passam os documentos de um arquivo, da sua produção à guarda permanente ou eliminação". São três essas fases: a produção, a utilização e a destinação, sendo que esta última pode destinar o documento para a guarda permanente ou eliminação.

A fase da produção é o momento em que os documentos são produzidos para cumprir sua função administrativa, sempre a partir da execução das atividades de um órgão. A fase de utilização abrange o momento em que o documento tramita e é utilizado dentro do órgão ou entidade, passando pelo momento de guarda até o tempo de destinação. A fase de destinação de documentos é aquela voltada ao fim que o documento terá após cumprir seu valor primário (leva em conta a sua utilidade para fins administrativos, legais e fiscais), podendo serem tais documentos recolhidos à guarda permanente em virtude de possuírem valor secundário (ou seja, valores diferentes daqueles para os quais foram criados) ou eliminados.

No caso de documentos digitais, o Arquivo Nacional (BRASIL, 2019, p. 50-51) define que "a melhor ferramenta para apoiar a gestão de documentos arquivísticos digitais é um sistema informatizado de gestão arquivística de documentos (SIGAD)". Corroborando, o Conarq (2015, p. 4) define que "os documentos arquivísticos digitais em fase corrente e intermediária devem, preferencialmente ser gerenciados por meio de um Sistema Informatizado de Gestão Arquivística de Documentos – SIGAD" e que, "a partir da destinação para a guarda permanente, ocorre uma alteração na cadeia de custódia passando a responsabilidade pela preservação dos documentos dos produtores para a instância de guarda".

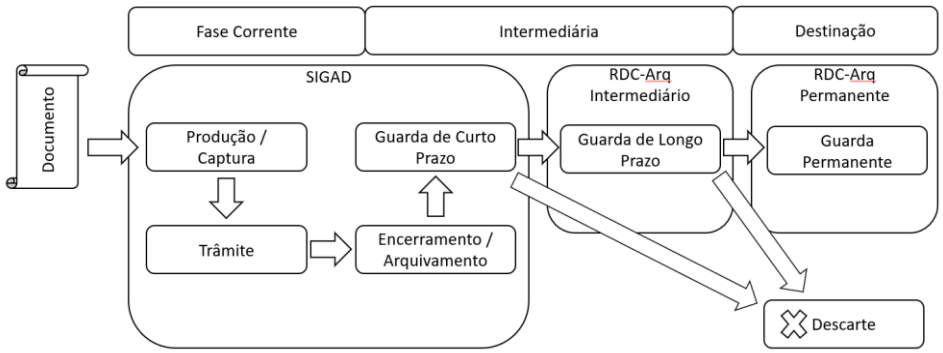
A obsolescência tecnológica é uma ameaça constante aos documentos digitas, pois coloca em risco a recuperação da cadeia de *bytes* da mídia de

armazenamento quando os leitores da mídia passam a ser escassos no mercado em virtude de uma nova tecnologia substituta, ou quando há escassez de sistemas capazes de decodificar a cadeia de *bytes* em informação legível ao usuário devido a adoção de um formato de arquivo mais novo. Neste sentido, o Conarq (2015, p. 3) observa que "a preservação de documentos arquivísticos digitais, nas fases corrente, intermediária e permanente, deve estar associada a um repositório digital confiável".

O Conarq também cunha o conceito de Repositórios Arquivísticos Digitais Confiáveis (RDC-Arq) para definir repositórios digitais confiáveis específicos para a preservação de documentos arquivísticos, e salienta que "um cenário que englobaria todo o ciclo de vida do documento seria o uso de um RDC-Arq para as idades corrente e intermediária, e de outro para a idade permanente" (CONARQ, 2015, p. 3).

Analisando as publicações do Conarq, podemos estabelecer que documentos em fase corrente e intermediária são geridos por um SIGAD e um RDC-Arq de nível intermediário. Quando estes documentos são destinados para a fase permanente, são transferidos para um RDC-Arq de nível permanente com uma alteração na cadeia de custódia. A figura 1 ilustra os processos arquivísticos relacionando aos sistemas envolvidos.

Figura 1 - Processos Arquivísticos e Sistemas Envolvidos



Fonte: Os autores.

A partir da figura 1, podemos visualizar quando o documento sai do SIGAD para o RDC-Arq, seja ele intermediário ou permanente. Neste momento, o documento fica vulnerável a adulterações, uma vez que, diferente da proteção dentro do SIGAD e do RDC-ARq, enquanto o documento está sendo transferido fica vulnerável à fragilidade e insegurança do canal, o qual precisa de aparatos e robustez de segurança.

Alguns recursos tecnológicos podem, e devem, ser implementados para sanar essa vulnerabilidade no canal de transmissão. O presente artigo apresenta alguns destes recursos que consideramos de suma importância para a garantia da integridade dos documentos arquivísticos.

4 FUNÇÕES CRIPTOGRÁFICAS DE HASH

Uma função criptográfica de *hash*, ou função de *hash* unidirecional, é uma função que, dado um fluxo de bites ou *bytes* de tamanho variável, sempre gera um resultado de tamanho fixo. Em outras palavras: um arquivo de computador "a", submetido a uma função *hash* "H", resulta em um resumo "r"; desta forma, temos que H(a)=r.

Além do tamanho fixo para o resumo, também chamado de *fingerprint* ou *digest*, uma função *hash* tem algumas outras caraterísticas importantes. A que se destaca, aqui, é o fato de ser uma função unidirecional, podendo ser entendida como uma função onde é simples calcular um resumo a partir de uma entrada, mas não é possível calcular a entrada a partir de um resumo.

Em uma função *hash*, a variação das entradas é infinita e a variação de resumos é finita, pois tem tamanho fixo, logo, sempre haverá colisões. Estas ocorrem quando um mesmo resumo é gerado a partir de arquivos de computador diferentes, ou seja, quando $H(a_1) = H(a_2)$. Considerando a existência de colisões, outra característica importante a ser destacada é a resistência à colisão. Segundo Ferguson e Schnneier (2003, p. 84, tradução nossa), "o requisito de resistência a colisão apenas afirma que embora colisões existam, elas não podem ser encontradas".

Dos algoritmos existentes de *hash*, destacaremos, neste trabalho, o MD5 e o SHA1 por serem referencias de segurança e eficiência.

> Entre os diversos algoritmos existentes, o *Message Digest* 4 (MD4) pode ser tido como um marco histórico, pois é nele que são baseados os algoritmos de *hash* atualmente mais populares, que compõem praticamente o padrão em relação à eficiência e segurança para a geração de *hashes*. Entre estes últimos tem-se: MD5 e SHA1. (SERAFIM, 2002, p. 40)

O algoritmo MD5 é uma extensão do algoritmo MD4 que, a partir de uma entrada de dados, produz um resumo com o tamanho de 128-bit. Boer e Bosselaers (1994) demonstraram fraquezas nesse algoritmo, as quais permitem a geração de colisões. No entanto, Schneier (1994) não acredita que existe impacto prático na segurança do algoritmo, mesmo considerando que existe, de fato, uma fraqueza na função de compressão.

O SHA é um padrão criado pelo *National Institute of Standards and Technology* (NIST) que especifica um conjunto de algoritmos para a produção de um resumo a partir de uma entrada de dados ou mensagem, objetivando detectar alterações na mensagem após a geração do resumo; o SHA-1 é apenas um dos algoritmos deste padrão, em que este estudo foca. De acordo com Serafim (2002), a diferença mais evidente entre o SHA-1 e o MD5 é que o resumo gerado pelo primeiro tem o tamanho de 160 bits, enquanto o do segundo é de 128 bits. Schneier (1996, p. 368) destaca que essa característica torna o SHA-1 mais resistente a ataques de força-bruta.

5 PADRÃO DE EMPACOTAMENTO BAGIT

O padrão Bagit é para empacotamento de arquivos de computador com a finalidade de guarda e transferência, criado pela *Library of Congrees* (Biblioteca do Congresso Americano) em parceria com a *California Digital Library* (Biblioteca Digital da Califórnia). O padrão não conhece a semântica do conteúdo por ele empacotado, permitindo, assim, a construção de pacotes com qualquer conteúdo digital.

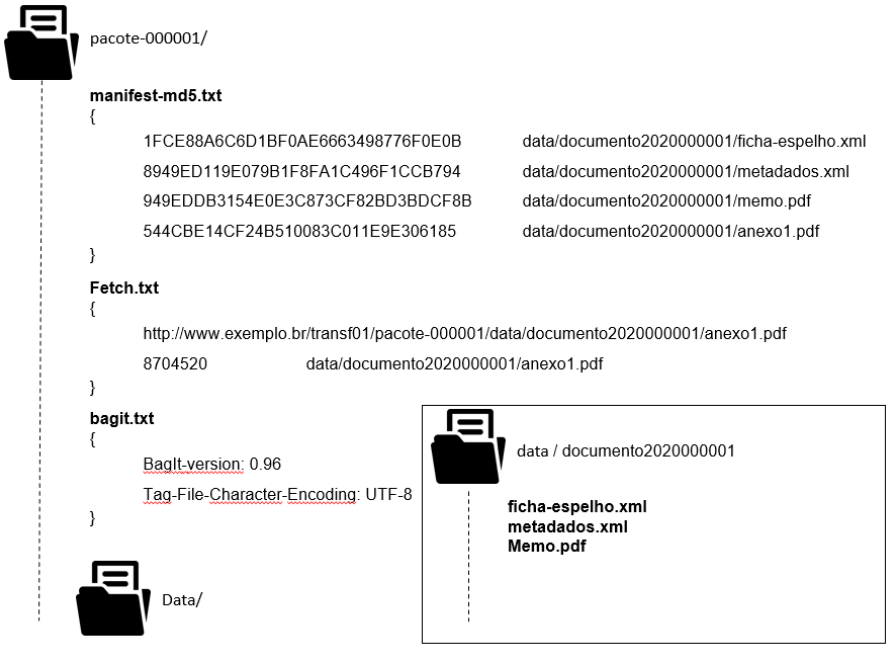
O pacote Bagit consiste em um diretório raiz contendo um arquivo de texto simples construído de forma a ser legível para um *software* e um subdiretório com os arquivos de conteúdo. O arquivo de texto contém o inventário de todos os arquivos do conteúdo empacotado, seguido por sua informação de soma de verificação.

Conforme a especificação do padrão, o arquivo com o inventário dos conteúdos empacotados é nomeado como manifest-algoritmo.txt, onde o texto "algoritmo" é substituído pelo texto referenciando o algoritmo de soma de verificação usado para a lista.

Para atender ao padrão, um último arquivo de texto, nomeado "baigit.txt", deve existir na pasta raiz do pacote. Este arquivo contém apenas duas linhas: a primeira com o texto "*BagIt-Version*: M.N", onde "M.N" é a versão do padrão usado, e a segunda linha com o texto "*Tag-File-Character-Encoding*: UTF-8", indicando o tipo de codificação binária usada para representar os caracteres.

O padrão prevê outros arquivos de texto dentro da pasta raiz do pacote e, para o escopo deste artigo, faz sentido destacar o "*fetch*.txt". Este arquivo contém, em uma lista, os arquivos de conteúdo que devem ser obtidos por *download* antes da análise de completude do pacote. Cada item da lista é representado por uma linha no arquivo contendo a URL do conteúdo, o tamanho do arquivo de conteúdo e o nome do arquivo de conteúdo incluído seu caminho relativo em relação ao diretório raiz.

Figura 2 - Exemplo de pacote de transferência no padrão Bagit



Fonte: Os autores.

Para ilustrar como a estrutura do padrão Bagit e, em específico, a funcionalidade do arquivo "*fetch*.txt", a figura 2 demonstra um pacote de transferência de documentos com um único documento composto por quatro componentes digitais. Neste exemplo, o arquivo "anexo1.pdf" seria muito grande para ser enviado junto com o pacote; desta forma, o arquivo "fetch.txt" define em qual endereço da *web* o arquivo "anexo1.pdf" pode ser adquirido, enquanto o arquivo "manifest-md5.txt" fornece a mensagem de integridade da informação de conteúdo do pacote completo.

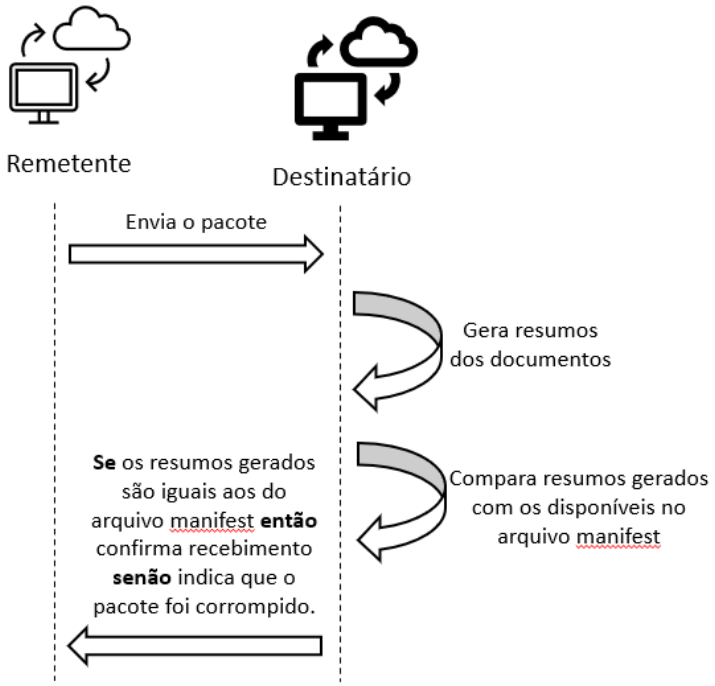
6 PROCEDIMENTOS PARA GARANTIA DA INTEGRIDADE DE DOCUMENTOS UTILIZANDO CÓDIGOS HASH

Os procedimentos propostos a seguir podem ser executados quando um conjunto de documentos é enviado do SIGAD para um RDC-Arq, visando garantir a integridade dos documentos enquanto não estão sob a gestão nem do SIGAD nem do RDC-Arq.

7 LISTAGEM DE RESUMOS HASH NO PACOTE DE TRANSFERÊNCIA

Este procedimento consiste na verificação da integridade dos documentos enviados em um pacote no formato Bagit a partir dos resumos *hash* enviados no arquivo *manifest* do pacote. A figura 3 ilustra esse procedimento.

Figura 3 - Listagem de resumos *hash* no pacote de transferência

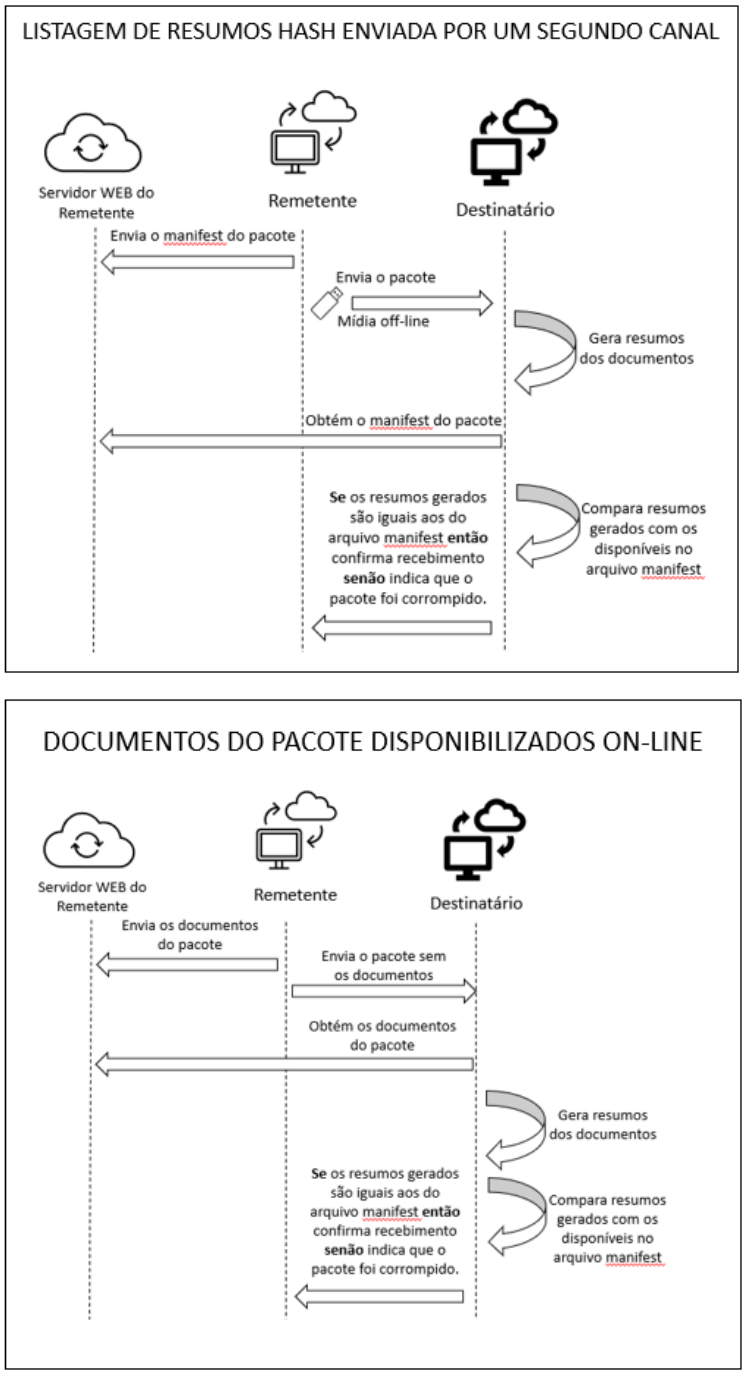


Fonte: Elaborado pelos autores.

8 LISTAGEM E RESUMOS HASH ENVIADA POR UM SEGUNDO CANAL

Este procedimento consiste na verificação da integridade dos documentos enviados em um pacote no formato Bagit a partir dos resumos *hash* enviados fora da estrutura de pasta do padrão Bagit, bem como por um canal diferente daquele usado para enviar o pacote original. A título de exemplo, o pacote de documentos poderia ser enviado em uma mídia off-line (um disco *blu-ray*, por exemplo) e o arquivo *manifest*, contendo os resumos *hash*, poderia ser disponibilizado pelo remetente, apenas para leitura, em um endereço *web*. O quadro esquerdo da figura 4 ilustra este procedimento.

Figura 4 - Listagem de resumos *hash* enviada por um segundo canal e documentos do pacote disponibilizados on-line



Fonte: Elaborado pelos autores.

9 DOCUMENTOS DO PACOTE DISPONIBILIZADOS ON-LINE

Este procedimento consiste no envio do pacote para o destinatário sem os documentos na pasta "DATA". O pacote deve conter, necessariamente, o arquivo *manifest* com os resumos *hash* dos documentos, e o arquivo "fetch.txt" contendo os *links* para o *download* dos mesmos documentos que o pacote deveria ter. O quadro à direita da figura 4 ilustra este procedimento.

10 CONCLUSÃO

Considerando que as funções criptográficas de *hash* são tidas como seguras por especialistas em criptografia para identificar alterações em documentos digitais, e que esta técnica já é considerada pelo padrão de empacotamento de documentos largamente utilizada para documentos arquivísticos, os procedimentos propostos para verificação de integridade visam garantir que alguém mal-intencionado, de posse de um pacote de documentos, não tenha acesso tanto aos documentos quanto aos resumos *hash*, impedindo, assim, a alteração intencional de um documento e a geração de um novo resumo.

O primeiro procedimento apresentado coloca em risco a integridade dos documentos, já que, ao longo de uma transferência, seria possível alterá-los e gerar uma nova lista de resumos a partir dos documentos modificados. Desta forma, o destinatário, ao verificar a integridade do pacote, não detectaria alterações.

O segundo e o terceiro procedimentos colocam um nível a mais na segurança das transferências, impedindo que alguém mal-intencionado tenha posse, ao mesmo tempo, dos documentos e dos resumos *hash*. Sendo assim, seria impossível alterar o conteúdo do pacote ao interceptá-lo sem que o destinatário identificasse a modificação.

A partir desta análise, podemos concluir que, durante as transferências entre SIGADS e RDC-Arqs, há a possibilidade de aumentar o nível de segurança na verificação de integridade dos documentos usando o segundo ou o terceiro procedimento proposto neste artigo.

REFERÊNCIAS

ABNT. **ABNT NBR 17572:2007** – Sistemas especiais de dados e informações – Modelo de referência para um sistema aberto de arquivamento e informação (SAAI). Rio de Janeiro: ABNT, 2007.

BOER, B.; BOSSELAERS, A. Collisions for the Compression Function of MD5. **Advances in Cryptology – EUROCRYPT'93**, Norway, p. 23-27, may 1993. Disponível em: https://link.springer.com/content/pdf/10.1007%2F3-540-48285-7.pdf. Acesso em: 28 ago. 2021.

BRASIL. **Decreto nº 10.278, de 18 de março de 2020**. Estabelece as diretrizes e bases da educação nacional. **Diário Oficial da União**, Brasília, DF, 2020.

BRASIL. **Decreto nº 8.539, de 8 de outubro de 2015**. Dispõe sobre o uso do meio eletrônico para a realização do processo administrativo no âmbito dos órgãos e das entidades da administração pública federal direta, autárquica e fundacional. **Diário Oficial da União**, Brasília, DF, 2015.

BRASIL. Ministério da Justiça. **Portaria Interministerial nº 1.677 de 7 de outubro de 2015**. Define os procedimentos gerais para o desenvolvimento das atividades de protocolo no âmbito dos órgãos e entidades da Administração Pública Federal. **Diário Oficial da União**, Brasília, DF, 2015.

CALIFORNIA DIGITA LIBRARY, USA; STANFORD LIBRARY, USA; LIBRARY OF CONGRESS, USA. **The BagIt File Packaging Format (v1.0), 2018**. Disponível em: https://datatracker.ietf.org/doc/html/rfc8493. Acesso em: 13 maio 2022.

CALIFORNIA DIGITA LIBRARY, USA; LIBRARY OF CONGRESS, USA. **The BagIt File Packaging Format (v0.97), draft-kunze-bagit-07.txt, 2012**. Disponível em: https://www.digitalpreservation.gov/documents/bagitspec.pdf. Acesso em: 13 maio2022.

CONSELHO NACIONAL DE ARQUIVOS. **Diretrizes para a Implementação de Repositórios Arquivísticos Digitais Confiáveis – RDC-Arq**. Rio de Janeiro: Arquivo Nacional, 2015.

CONSELHO NACIONAL DE ARQUIVOS. **E-ARQ BRASIL**: Modelo de Requisitos para Sistemas Informatizados de Gestão Arquivística de Documentos. Rio de Janeiro: Arquivo Nacional, 2011.

Digital criteria standard for eletronic record managment software applications: DOD 5015.2-STD, 2002.

FERGUSON, N. F.; SCHNEIER, B. **Practical Criptography**. 1. ed. Wiley, Indiana, USA: Wiley, 2003.

NIST. **Secure Hash Standard**. FIPS PUB 180-1. Washington D.C. 1995.

PREMIS EDITORIAL COMMITTEE. **Data Dictionary for Preservation Metadata**: PREMIS version 3.0, 2015. Disponível em: http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf. Acesso em: 11 maio 2022.

RIVEST, R. **The MD4 Message-Digest Algorithm**: RFC 1320. 1992.

RIVEST, R. **The MD5 Message-Digest Algorithm**: RFC 1321. 1992.

SCHNEIER, B. **Applied Cryptography**. 2. ed. New York: John Wiley & Sons, 1996.

SERAFIM, V. da S. **Um verificador seguro de integridade de arquivos**. 2012. Disponível em: http://hdl.handle.net/10183/2756. Acesso em: 19 maio 2021.

THE CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS (CCSDS). **Recommendation for Space Data System Practices** – Reference model for an open archival information system (OAIS) – CCSDS 650.0-M-2. Washington, DC, USA. Disponível em: https://public.ccsds.org/pubs/650x0m2.pdf. Acesso em: 11 maio 2022.

THE CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS (CCSDS). **Audit and certification of trustworthy digital repositories**. Washington, DC, USA. Disponível em: https://public.ccsds.org/Pubs/652x0m1.pdf. Acesso em: 11 maio 2022.