Avaliação Experimental da Plataforma Archive Fixity Anchor em Cenários de Adulteração de Dados em Ambiente Simulado de Repositório Digital Confiável

Filipy Galiza e Rostand Costa

Programa de Pós-Graduação em Informática (PPGI), Laboratório de Aplicações de Vídeo Digital (LAVID)
Centro de Informática – Universidade Federal da Paraíba (UFPB)
filipy@uepb.edu.br, rostand@lavid.ufpb.br

Resumo: Para auxiliar os Repositórios Digitais Confiáveis (RDC) em sua missão de garantir a fixidez dos documentos arquivísticos digitais, uma plataforma denominada Archive Fixity Anchor pode ser utilizada concomitantemente para que, a partir do uso combinado de Árvores de Merkle e *blockchains*, a Plataforma possa fornecer mecanismos confiáveis para o registro e auditoria da fixidez dos acervos nos repositórios. Este trabalho visa documentar a avaliação do comportamento desta plataforma e de um software de repositório no tratamento da fixidez de um acervo de objetos digitais submetidos a cenários simulados de adulteração de dados.

Palavras-chave: *preservação digital; fixidez; RDC; blockchain*.

1. Introdução

Sistemas de *software* tidos como referência para construção de repositórios digitais [1], como o Archivematica [2] e o RODA [3] executam a verificação da fixidez (*fixity check*) dos objetos digitais sob seus domínios com base em procedimentos padrão: comparando informações de fixidez (*fixity information*) geradas instantaneamente a partir dos objetos com *valores* de referência previamente registrados localmente referentes a esses objetos.

Esses processos de verificação normalmente utilizados podem ser insuficientes para garantir a fixidez em um Repositório Digital Confiável (RDC) frente as ameaças que as informações digitais estão expostas, principalmente no longo prazo, devido a degradação ocasional [4] ou adulteração intencional dos dados [5]. Portanto, a vulnerabilidade dos valores de referência para o controle de fixidez põe em risco a confiabilidade do acervo preservado, pois a desconfiança dessas informações pode fazer com que um acervo adulterado seja considerado legítimo, assim como a confiança em um acervo legítimo pode ser perdida.

Diante essa preocupação, Galiza e Costa [6] propõem uma abordagem baseada no uso de Árvores de Merkle e *blockchains* para que sejam gerados e armazenados imutavelmente valores de referência para o controle de fixidez dos acervos em RDC, além dos recursos de redundância utilizados pelos repositórios. A partir dessa proposta foi desenvolvida uma plataforma denominada Archive Fixity Anchor (AFA), que implementa os principais mecanismos para apoiar os conceitos defendidos pelos autores.

Os códigos e informações adicionais acerca da Plataforma desenvolvida podem ser encontrados em [7] e no seu atual estágio de desenvolvimento permite registrar manualmente os Pacotes de Arquivamento de Informações (Archival Information Packages – AIP) de um acervo digital com seus identificadores associados e viabiliza a auditoria da fixidez desses AIP a partir da consulta de valores confiáveis de referência ancorados em *blockchain*.

Este trabalho visa apresentar os resultados obtidos a partir da avaliação experimental da Plataforma em um ambiente simulado de RDC, montado utilizando os elementos mínimos necessários a essa tarefa, resumido no uso da AFA e do Archivematica como *software* de repositório digital, assim como todo o arcabouço tecnológico necessário a correta execução desses elementos.

Como avaliação experimental foram simulados três possíveis cenários considerados passíveis de acontecimento na prática de um repositório digital, onde se tenta simular algum tipo de alteração maliciosa no acervo, visando em todos os cenários avaliar o comportamento do *software* de repositório Archivematica e da plataforma AFA no âmbito da verificação de fixidez.

A primeira seção deste trabalho realiza uma contextualização acerca da temática abordada; na segunda seção se apresenta resumidamente a metodologia adotada na execução da avaliação; na seção 3 são apresentados os experimentos realizados e os resultados obtidos; e na seção 4 são feitas discussões e considerações finais.

2. Metodologia

Para a execução da avaliação, um ambiente de repositório digital foi simulado utilizando o *software* Archivematica, escolhido como *software* de repositório de acordo com as recomendações de Rodrigues [1] para tal. Como esperado, também compõe o ambiente o *software* AFA, alvo da avaliação deste trabalho e que visa apoiar as ações relacionadas a fixidez nesse repositório.

Esse ambiente de experimentos foi montado seguindo as recomendações de *hardware* e *software* necessários para a correta execução dos sistemas escolhidos de acordo com suas documentações [8][9], seguindo a instalação e configuração padrão ou recomendada para os sistemas e que resultou em um ambiente de repositório funcional e apto a ingestão de objetos para preservação.

Além da correta execução dos sistemas, se faz necessário o registro de objetos digitais no Archivematica para que esses sejam transformados em pacotes aptos para a preservação (AIP).

Como objetos digitais, foram selecionados documentos de imagem disponíveis em [7], que são registrados no Archivematica de acordo com as instruções documentadas em [8] na medida em que se executa os cenários de simulação.

Após os registros, os pacotes resultantes podem ser consultados pela aba “Packages” do Archivematica Storage Service, onde são apresentadas informações como o identificador único (UUID) e o local atual de armazenamento (caminho no sistema) para se ter acesso ao pacote, informações necessárias ao registro do mesmo na AFA, realizados de acordo com as instruções documentadas em [7] no decorrer da execução dos cenários de simulação.

O tratamento dos objetos e sua transformação em AIP são conduzidos automaticamente pelo Archivematica que, por padrão, permite o acompanhamento e possíveis intervenções pelo usuário, dentre as quais estão algumas opções de personalização para adequação aos padrões e objetivos de preservação do repositório. Mas para o registro dos objetos aqui relatados, além de não adicionar metadados durante esses processos iniciais, tais opções de intervenções foram definidas com valores recomendados pela documentação do Archivematica ou, na ausência de recomendação, definidas com as opções que resultassem em menos etapas de processamento.

Em cada um dos cenários de avaliação acontecem os processos de registro de objetos, a verificação da fixidez desses, a adulteração de dados e posterior repetição da conferência da fixidez, para, com isso, observar o comportamento dos *softwares* com relação a identificação de inconsistências nos AIP nos cenários avaliados.

Se espera que a AFA consiga identificar inconsistências de fixidez mesmo nos cenários em que o Archivematica possa não identificar, podendo demonstrar seu potencial incremento nas ações de preservação dos repositórios no aspecto de verificação e garantia da fixidez dos acervos.

Para a verificação da fixidez pelo Archivematica será utilizado um mecanismo nativo para esse controle, que acompanha o Archivematica Storage Service e que está disponível através de um API *endpoint* com funcionamento e uso documentado em [9]. Para facilitar sua referência no texto, esse método de verificação será referenciado apenas como *fixity endpoint*.

Na AFA, a verificação de fixidez acontece através do recurso de auditoria e seu uso está documentado em [7].

3. Avaliação e Resultados

Os três cenários de avaliação experimental e os resultados encontrados a partir deles são documentados nesta seção. Neles, foram apenas simuladas adulterações consideradas maliciosas e não desejadas para acontecimento em um ambiente de RDC.

3.1 Cenário 1: Adulteração simples de um pacote de informação

Nesse cenário se visou avaliar o comportamento do Archivematica e da AFA frente a simulação de uma possível adulteração de pacotes de informação, que poderia acontecer de forma maliciosa ou ocasional.

Para a simulação deste cenário foram registrados quatro objetos digitais no Archivematica, o que resultou na criação de quatro AIP, os quais foram devidamente registrados na AFA.

Após a confirmação do estado consistente da fixidez em ambos os sistemas, um dos pacotes foi escolhido aleatoriamente e teve seu conteúdo intencionalmente modificado, mas foi adulterado de forma a permanecer com as mesmas informações do pacote original numa tentativa de passá-lo como autêntico e inalterado nos processos de verificação de fixidez, simulando uma corrupção maliciosa.

Após efetivado a alteração no pacote, uma nova checagem de fixidez foi requisitada através da consulta ao *fixity endpoint* e da auditoria na AFA para se obter o estado atualizado de fixidez desse pacote, sabendo que esse não corresponde mais ao seu estado original.

A ferramenta do Archivematica detectou a inconsistência do pacote e informou sobre o estado de falha da fixidez desse, ainda fornecendo algumas informações adicionais sobre o problema (relatando quantos arquivos e o tamanho total em *bytes* esperado e o que foi constatado).

De maneira semelhante, a verificação de fixidez do pacote na AFA também alertou sobre a inconsistência de sua fixidez, mas diferente do Archivematica, a AFA não fornece detalhes sobre a inconsistência do registro. Isso acontece devido as diferentes formas de implementação do controle da fixidez, que no Archivematica se baseia no uso das especificações do formato Bagit [10] para guiar a formação dos pacotes, o que inclui informações adicionais sobre seu conteúdo, enquanto a AFA abstrai as informações do conteúdo e apenas considera as informações de *hash*.

Entretanto, independente dos detalhes fornecidos, neste cenário, tanto o Archivematica quanto a AFA alertaram corretamente para a falha da fixidez do pacote adulterado, ficando o repositório munido de duas ferramentas que lhe garante a correta detecção de corrupção de um pacote, como o simulado nesse cenário.

3.2 Cenário 2: Adulteração de um pacote de informação e de suas respectivas informações de fixidez

Nesse cenário se visou avaliar o comportamento do Archivematica e da AFA frente a simulação de uma possível adulteração puramente maliciosa de um pacote de informação, onde um atacante tentaria forjar a autenticidade de um pacote alterado.

Para a simulação deste cenário foram registrados outros quatro objetos digitais no Archivematica, o que resultou na criação de mais quatro AIP, cujos foram devidamente registrados na AFA.

Após a confirmação do estado consistente da fixidez em ambos os sistemas, um dos pacotes foi escolhido aleatoriamente e teve seu conteúdo intencionalmente modificado, mas foi adulterado de forma a permanecer com as mesmas informações do pacote original numa tentativa de passá-lo como autêntico e inalterado nos processos de verificação de fixidez, simulando uma corrupção maliciosa.

Além disso, nesse cenário o pacote foi reconstruído de acordo com as especificações do Bagit, com suas novas informações de fixidez com auxílio da ferramenta bagit-python, almejando o reconhecimento do pacote como autêntico pelo *fixity endpoint* do Archivematica.

Após a manipulação do pacote, uma nova checagem de fixidez foi requisitada através da consulta ao *fixity endpoint* e da auditoria na AFA para se obter o estado atualizado de fixidez desse pacote, sabendo que esse não corresponde mais ao seu estado original.

Nesse caso, a ferramenta do Archivematica considerou o pacote em questão como se permanecesse com sua fixidez consistente, ainda que o conteúdo do pacote estivesse alterado. Esse falso sucesso na verificação é mantido ao se obter detalhes no sistema sobre a verificação de fixidez do pacote, levando o usuário a crer que a fixidez do pacote continua inabalada.

Por outro lado, a verificação de fixidez do pacote na AFA retornou um alerta sobre a inconsistência de sua fixidez.

Essa divergência de resultados acontece devido as diferenças na implementação do controle da fixidez, como já previamente esclarecido e, nesse caso, mostrou a relevância do uso da AFA para apoiar o controle de fixidez em um repositório, que apenas com a ferramenta de verificação nativa do Archivematica pôde ser facilmente burlado nesse cenário.

3.3. Cenário 3: Adulteração de um pacote de informação e de suas respectivas informações de fixidez registradas na AFA

Nesse cenário se visou avaliar o comportamento da AFA frente a simulação de uma possível adulteração puramente maliciosa de um pacote de informação e das informações na AFA relacionadas a fixidez desse pacote, onde um atacante tentaria forjar a autenticidade de um pacote alterado manipulando os dados da AFA para que essa apoie sua ação.

Para esse cenário se optou por utilizar o mesmo conjunto de pacotes já previamente registrados nos sistemas no Cenário 2, dando continuidade aos processos iniciados nesse cenário anterior e eliminando a necessidade de refazer os mesmos processos a fim de se alcançar o estado no qual se findou o experimento documentado nesse último cenário.

Foi considerada, então, a escolha do mesmo pacote alvo do cenário anterior, onde esse foi devidamente manipulado e se conheceu o comportamento do *fixity endpoint* do Archivematica e da AFA frente a essa manipulação. Mas agora se pretende forçar a situação para que a AFA também apoie a legitimidade desse pacote adulterado.

Conhecendo o funcionamento da Plataforma, foi possível adulterar seus dados locais relacionados ao pacote alvo para que esse possa ter sua fixidez considerada consistente pela AFA.

Para a manipulação dos dados bastou estruturar as informações dos pacotes na mesma ordem em que se foram executados os registros legítimos, posicionando devidamente os novos valores maliciosos gerados a partir de uma nova Árvore de Merkle. No entanto, apesar da devida manipulação, ao se chamar uma auditoria na plataforma, essa continuou a alertar para a inconsistência nos dados relacionados a fixidez do pacote alvo.

Essa detecção de inconsistência é sentida pela plataforma devido à divergência entre as informações de fixidez (*hash* raiz das Árvores) armazenadas localmente e de seu valor âncora salvo na rede *blockchain*. Mas, ainda que se tente alterar apenas o valor *hash* raiz local para tentar coincidir com o valor âncora, a plataforma detecta que esse valor não condiz com o valor *hash* raiz gerado pela Árvore de Merkle que considere o *hash* do pacote adulterado e continua alertando sobre a inconsistência de fixidez dos dados relacionados ao pacote adulterado.

4. Discussões e Considerações Finais

Este trabalho se propôs a documentar a realização da avaliação experimental da plataforma desenvolvida Archive Fixity Anchor em cenários de adulteração de dados em ambiente simulado de RDC utilizando o *software* Archivematica.

A partir dos experimentos aqui documentados pôde se observar que a plataforma AFA se comportou de acordo com o esperado para sua versão implementada e satisfez a expectativa dos resultados: alertando sobre o estado genuíno de fixidez daqueles pacotes em si registrados e de suas próprias informações locais sobre os registros a partir do uso de um valor âncora confiável em *blockchain*, mesmo quando as inconsistências foram despercebidas pelo mecanismo de controle do Archivematica.

Se considera que uma das formas que um atacante pode ter para tentar contornar o correto resultado entregue pela Plataforma, a exemplo do resultado obtido no Cenário 3, seria o atacante descobrindo uma outra transação na mesma rede *blockchain* que carregue dados que represente o valor âncora necessário para fazer com que os dados locais adulterados possam parecer autênticos e, com isso, ele poderia ter um endereço de transação coerente para substituir nos dados locais da Plataforma e finalmente validar os registros adulterados. Porém, essa é uma condição improvável de se alcançar, principalmente se for considerado que em uma implementação ideal, a Plataforma prevê o uso de redes e algoritmos de *hash* redundantes, o que reduziria ainda mais a probabilidade do atacante coincidir todas essas variáveis a seu favor. Isso tudo sem considerar a marca temporal das transações, que a Plataforma não considera em suas análises de fixidez, mas que se vier a ser considerada de alguma forma, se torna mais uma variável a pesar na obstrução do ataque.

Uma outra forma mais simples de burlar o mecanismo de auditoria da AFA seria executar os registros normalmente informações derivadas de adulteração, obtendo um registro legítimo para informações adulteradas e a partir disso, tentar fazer um registro local se passar por outro. Esse ataque se mostra facilmente viável principalmente quando o registro alvo do ataque é relativamente recente, uma vez que a marca temporal pode não ser tão decisiva nesses casos. Porém, nesse último caso pode se pôr em discussão o interesse de ataques a registros recentes e para todos os casos pode se considerar a implementação de recursos que tratem as transações realizadas pela carteira do repositório na rede *blockchain* com fins de registro como referência para as transações registradas localmente, além da possibilidade de se considerar a marca temporal das transações como mais um elemento de controle.

Também se estima que uma outra forma de enganar o correto funcionamento da Plataforma e não considerada nos experimentos poderia ser a de adulteração de seu código fonte a fim de direcionar seu funcionamento em favor de um ataque. Para esses casos resta a realização de uma auditoria no código fonte da Plataforma em execução ou a obtenção, sempre que se considerar pertinente, dos códigos oficiais da Plataforma nos casos em que não se trabalha com versões modificadas dessa.

O fato da realização dos experimentos utilizando um número reduzido de amostras se deu pela necessidade de intervenção manual para registro na AFA e a racionalização do tempo desta pesquisa.

Apesar de não abordado nesses experimentos, o Archivematica dispõe de mecanismos para automatizar o registro de objetos digitais, mas a indisponibilidade desse tipo de recurso na AFA compromete atualmente o registro de grandes acervos (centenas ou milhares de pacotes) e consequentemente seu uso em produção.

Bibliografia

[1] Rodrigues, M.M. (2015) Repositório Arquivístico Digital Confiável para o Patrimônio Documental Oriundo do Processo Judicial Eletrônico. Dissertação de Mestrado. Programa de Pós-Graduação em Patrimônio Cultural. Universidade Federal de Santa Maria. Online: https://repositorio.ufsm.br/handle/1/11050.

[2] Archivematica: open-source digital preservation system. Online: https://www.archivematica.org/. Acesso em 18/01/2021.

[3] RODA - Repositório para preservação de informação digital - KEEP SOLUTIONS. Online: https://www.keep.pt/produtos/roda-repositorio-para-preservacao-de-informacao-digital/. Acesso em 18/01/2021.

[4] Wright, R.; Miller, A.; Addis, M. (2009) The Significance of Storage in the “Cost of Risk” of Digital Preservation. *International Journal of Digital Curation* 4(3): 104–122. DOI: 10.2218/ijdc.v4i3.125

[5] National Research Council (2005) Building an Electronic Records Archive at the National Archives and Records Administration: Recommendations for a Long-Term Strategy. National Academies Press, Washington, D.C. DOI: 10.17226/11332

[6] Galiza, F.; Costa, R. (2020). Uma Abordagem Baseada em DLTs para Garantia da Fixidez de Repositórios Digitais Confiáveis (RDCs). Anais do III Workshop em Blockchain Teoria, Tecnologias e Aplicações, pp. 41–54. SBC. DOI: 10.5753/wblockchain.2020.12432

[7] Plataforma para auxiliar os Repositórios Digitais Confiáveis a garantir a fixidez de seus acervos. Online: https://bit.ly/393zvol. Acesso em 22/07/2021.

[8] Archivematica Quick-Start Guide | Documentação (Archivematica 1.12.1) | Archivematica: open-source digital preservation system. Online: https://www.archivematica.org/pt-br/docs/archivematica-1.12/getting-started/quick-start/quick-start/. Acesso em 18/01/2021.

[9] Fixity | Documentation (Archivematica Storage Service 0.17.1) | Archivematica: open-source digital preservation system. Online: https://www.archivematica.org/en/docs/storage-service-0.17/fixity/. Acesso em 22/07/2021.

[10] The BagIt File Packaging Format (V1.0). Online: https://tools.ietf.org/html/rfc8493. Acesso em 18/01/2021.