



ICICT

Instituto de Comunicação e Informação
Científica e Tecnológica em Saúde



PLANO DE PRESERVAÇÃO DIGITAL

Repositório Institucional Arca

(Versão 1.0)



arca

2020



Ministério da Saúde

FIOCRUZ

Fundação Oswaldo Cruz

EXPEDIENTE

Fundação Oswaldo Cruz - Fiocruz
Presidente Nísia Trindade Lima

Vice-Presidência de Ensino, Informação e Comunicação - VPEIC
Cristiani Vieira Machado

Instituto de Comunicação, Informação Científica e Tecnológica em Saúde - ICICT
Rodrigo Murtinho de Martinez Torres

Vice-Direção de Informação e Comunicação - ICICT
Tânia Cristina Pereira dos Santos

Centro de Tecnologia de Informação e Comunicação - CTIC
Aldo Lucio Pontes Moura

Seção de Informação
Luciana Danielli de Araujo

Coordenação do Arca - Repositório Institucional
Claudete Fernandes Queiroz

Autoria, Produção e Revisão do Texto
Andréa Ferreira Gonçalves do Nascimento

Colaboração:
Claudete Fernandes de Queiroz
Luciana Danielli de Araujo
Éder de Almeida Freire
Raphael Belchior Rodrigues
Angelo José Moreira Silva
Lucas Zinato Carraro

Programação Visual - Multimeios
Valéria de Sá

Catálogo-na-Publicação (CIP) - Brasil
Seção de Informação CTIC/ICICT

N244c

Nascimento, Andréa Ferreira Gonçalves do.
Plano de Preservação Digital: Arca - Repositório Institucional / Andrea Ferreira
Gonçalves do Nascimento. -- Rio de Janeiro, Fundação Oswaldo Cruz/ICICT, 2020.
20 p. : tab.

1. Preservação Digital - Plano. 2. Arca - Repositório Institucional. 3. Acesso
Aberto. I. Fundação Oswaldo Cruz. Instituto de Comunicação, Informação
Científica e Tecnológica em Saúde. II. Título.

CDD 025.8

Bibliotecário responsável: Luciana Danielli de Araujo - CRB-7 5024



Repositório Institucional da Fiocruz

PLANO DE PRESERVAÇÃO DIGITAL
REPOSITÓRIO INSTITUCIONAL ARCA
(VERSÃO 1.0)

2020

© 2020 Fundação Oswaldo Cruz. Instituto de Comunicação, Informação Científica e Tecnológica em Saúde. Todos os direitos reservados.

É permitida a reprodução parcial ou total desta obra, desde que citada a fonte e que não seja para venda ou qualquer outro fim comercial.

Este documento está disponível em Acesso Aberto de acordo com a Política de Acesso Aberto ao Conhecimento da Fiocruz.

Agradecimento especial ao Grupo de Trabalho sobre Preservação Digital do Projeto Preservo da Fundação Oswaldo Cruz. Vice-Presidência de Ensino, Informação e Comunicação.

Elaboração, distribuição e informações:

Fundação Oswaldo Cruz.

Instituto de Comunicação, Informação Científica e Tecnológica em Saúde. Prédio do Multimeios
Av. Brasil, 4365 - Manguinhos - CEP 21045-900 - Rio de Janeiro, RJ, Brasil.

Tel. (21) 2598-4242. <http://www.icict.fiocruz.br>.

SUMÁRIO

1. Introdução	7
2. Cenário institucional	7
2.1 Estrutura organizacional	7
2.2 Políticas de preservação relacionadas	7
2.3 Obrigações legais	7
3. Descrição do acervo	8
3.1. Identificação e escopo	8
3.2. Aspectos organizacionais	10
3.3. Aspectos técnicos	10
3.4. Requisitos dos usuários	10
3.5. Práticas correntes de preservação	11
4. Avaliação de riscos	12
5. Planejamento de obsolescência	14
5.1. Requisitos para preservação	14
5.2. Estratégias de preservação avaliadas	14
5.3. Seleção da estratégia	16
5.3.1. Fluxograma de trabalho	16
6. Recuperação de desastres	18
7. Revisão do Plano de Preservação Digital	18
8. Documentos relacionados	19

1. INTRODUÇÃO

O presente Plano de Preservação Digital analisa potenciais estratégias de preservação digital e apresenta as orientações e procedimentos operacionais para a preservação a longo prazo do acervo digital depositado no Repositório Arca, de acordo com as diretrizes estabelecidas no Programa de Preservação Digital da Fiocruz.

2. CENÁRIO INSTITUCIONAL

2.1 Estrutura organizacional

O Arca – Repositório Institucional da Fiocruz, lançado oficialmente em abril de 2011, tem como missão reunir, hospedar, *preservar*, disponibilizar e dar visibilidade à produção intelectual da Instituição.

O Instituto de Comunicação e Informação Científica em Saúde (Icict/Fiocruz) coordena o Comitê Gestor do Arca e é responsável pelo seu desenvolvimento tecnológico e metodológico, bem como pela operação e manutenção do repositório, de forma compartilhada com as demais unidades da Fiocruz, através dos Núcleos de Acesso Aberto ao Conhecimento (NAACs).

A gestão financeira do Arca é feita com recursos da Fiocruz especificamente alocados para esta finalidade, conforme estabelecido na Política de Acesso Aberto ao Conhecimento.

2.2 Políticas de preservação relacionadas

A [Política de Acesso Aberto ao Conhecimento](#), aprovada pela Fiocruz em 2014, institui o Repositório Institucional Arca como principal instrumento de realização do acesso aberto na instituição. A Política torna mandatório o depósito no Repositório Institucional Arca das dissertações e teses defendidas nos Programas de Pós-graduação da Fiocruz e dos artigos científicos de seus colaboradores publicados em periódicos, respeitados os períodos de embargo dessas publicações.

Em 2018, a Fiocruz lançou a [Política de Preservação dos Acervos Científicos e Culturais](#), que inclui aspectos voltados à preservação digital, debatidos com maior detalhe pelo Grupo de Trabalho para desenvolvimento do Programa de Preservação Digital da Fiocruz, do qual participa o Repositório Institucional Arca.

2.3 Obrigações legais

O acesso às obras intelectuais depositadas no Repositório Institucional Arca deve observar as hipóteses legais de sigilo obrigatório vigentes, em especial quanto aos direitos e obrigações relativos à propriedade industrial, proteção da propriedade intelectual de programa de computador, inovação e pesquisa científica e tecnológica, e acesso a informações, bem como períodos de embargo estabelecidos pelas publicações.

Para tal, utilizamos “Termos de Cessão” que são documentos autorizados pelos autores visando a cessão dos direitos no RI. Esses termos estão disponíveis na Intranet através do link <https://intranet.fiocruz.br/intranet> – Acesso Aberto – Governança Institucional – Acesso Aberto Fiocruz – Documentos e termos de cessão.

3. DESCRIÇÃO DO ACERVO

3.1 Identificação e escopo

O acervo do Arca é composto por documentos de caráter técnico-científico produzidos por profissionais ou órgãos vinculados à Fiocruz. O acervo do Arca está organizado em comunidades que correspondem às unidades técnico-científicas da Fiocruz. Cada comunidade pode reunir os seus documentos em subcomunidades, que representam os programas de pós-graduação, e/ou em diferentes coleções, que reúnem os documentos por tipologia¹.

Em dezembro de 2019, o acervo do Arca contabilizou um total de 34.953 registros, dos quais 82% correspondem a documentos de depósito mandatório – artigos de periódico (71%), dissertações (20,5%), teses (8%) e preprints (0,5%). A previsão de crescimento anual do acervo é de 30% sobre o total de registros.

As tipologias de documento disponibilizadas nas coleções do Arca com os respectivos formatos de arquivo aceitos para depósito e acesso², de acordo com as diretrizes do [Manual de Digitalização](#) e do [Plano Operativo do Arca](#), são apresentadas na Tabela 1.

Tabela 1 - Formatos de arquivo aceitos para depósito no Arca, segundo a tipologia do documento

Tipologia de documento	Formato de arquivo aceito
anais de congresso	PDF ou PDF/A
artigos de periódicos	
capítulos de livros	
cartas	
dados de pesquisa	
dissertações de mestrado	
livros publicados	
manuais e procedimentos técnicos	
patentes	
preprint	
relatórios de pesquisa	
relatórios institucionais	
revistas/boletins	
software	
termos de referência	
teses de doutorado	
trabalhos apresentados em eventos	
trabalhos de conclusão de curso	
áudio	MP3
fotografias	JPEG ou PNG
jogos	RAR
recursos educacionais	DOC, ODT, MP4, PDF, ZIP
softwares	APP, EXE, ZIP
vídeos	MP4

Fonte: Manual de digitalização da Fiocruz (2018) e Plano Operativo Arca (2017)

¹ A lista de comunidades, subcomunidades e coleções do Arca está disponível em: <https://www.arca.fiocruz.br/community-list>

² Por padrão, o software Dspace, onde opera o Arca, aceita os seguintes formatos para depósito, porém nem todos são utilizados atualmente no Arca: pdf, xml, txt, asc, htm, html, css, doc, docx, ppt, pptx, xls, xlsx, jpeg, jpg, gif, png, tiff, tif, aiff, aif, aife, au, snd, mp3, wav, mpeg, mpg, mpe, rtf, vsd, fm, bmp, psd, pdd, ps, eps, ai, mov, qt, mpa, abs, mpega, mpp, mpx, mpd, ma, látex, tex, dvi, sgm, sgml, wpd, ra, ram, pcd, ott, odt, oth, odm, odg, otg, odp, otp, ods, ots, odc, odf, odb, odi, oxt, sxw, stw, sxc, stc, sxd, std, sxi, sti, sxc, sxm, sdw, sgl, sdc, sda, sdd, sdp, smf, sds, sdm, rdf, mp4, epub, zip.

O quantitativo de arquivos depositados por formato é apresentado na Tabela 2.

Tabela 2 - Formatos de arquivo depositados no Arca e quantitativo (Novembro/2019)

Formato	Total
DOC	27
DOCX	98
EPUB	153
FMP3	1
GIF	3
HTML	7
JPEG	5745
MOV	1
MP3	13
MP4	898
MPEG	1
ODG	2
ODP	5
ODS	1
ODT	4
PDF	34423
PNG	99
PPT	53
PPTX	48
TIFF	6
TXT	71383
XLS	14
XLSX	10
XML	72
ZIP	19

Fonte: Arca (2019)

Todos os objetos digitais depositados no Arca possuem metadados descritivos associados, baseados no padrão Dublin Core (DC). O conjunto mínimo de metadados descritivos adotados para cada tipologia de documento está descrito no Manual de tratamento de dados do Arca³. Atualmente, o Arca não contempla o uso de metadados de preservação.

Para o autoarquivamento de *artigos científicos* diretamente pelos autores, o sistema requer o preenchimento mínimo de 3 campos de metadados: Autor, Título e Ano da publicação, além do upload do documento. Para as demais tipologias de documento, devem ser preenchidos todos os metadados obrigatórios, conforme estabelecido no [Manual de Tratamento de Dados](#) do Arca. Após completado o autoarquivamento, o bibliotecário responsável pela unidade de origem do pesquisador complementa os metadados necessários, certifica o registro e o torna disponível para consulta. O autor receberá uma notificação por e-mail assim que o depósito se tornar parte da coleção ou se, por alguma razão, houver problemas com o envio. O autor também pode verificar o status de seu depósito em sua "Área Pessoal" no Arca.

³ Manual de tratamento de dados: preenchimento de metadados para entrada no Arca - Repositório Institucional da Fiocruz. 3. ed. rev. e aum. Disponível em: <https://www.arca.fiocruz.br/handle/icict/13257>.

Caso o documento seja depositado pelos profissionais do NAAC ou das bibliotecas de referência, ou migrado a partir de outro sistema, os metadados descritivos são preenchidos da forma mais completa possível.

3.2 Aspectos organizacionais

A equipe de gestão operacional do Arca é composta por 11 profissionais vinculados ao Icict, sendo: 2 bibliotecárias coordenadoras/gestoras, 2 bibliotecários bolsistas, 3 analistas de sistemas, 3 técnicos de biblioteca e 1 estagiária.

A gestão operativa do Arca é realizada pela equipe, em conjunto com os NAACs de cada unidade da Fiocruz. Os fluxos operacionais e os mapas de processos que descrevem as diferentes modalidades de entrada de dados no repositório estão descritos no [Plano Operativo do Arca](#).

3.3 Aspectos técnicos

O Arca opera no sistema DSpace, atualmente na versão 4.7, com migração prevista para a versão 6.3. O DSpace é um software livre de código aberto, que permite o gerenciamento da produção científica em meio digital e em diversos formatos, e utiliza o protocolo Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) para interoperabilidade com outros sistemas e recuperação por sistemas de busca na web.

A estrutura tecnológica que suporta o Arca é composta por 3 instâncias, localizadas no espaço de armazenamento da Nuvem Fiocruz: (1) servidor de produção, onde estão instalados os softwares DSpace, Jetty, PostgreSQL e Java; (2) servidor de homologação, que é uma réplica do servidor de produção e é utilizado para testes; e (3) servidor de treinamento.

Atualmente, os sistemas e conteúdo do Arca ocupam aproximadamente 1 TB de espaço e tem uma previsão de crescimento de 30% ao ano sobre o total de registros na base.

3.4 Requisitos dos usuários

A equipe de gestão operacional do Arca desempenha o papel de Administradores do sistema, com acesso total à gestão dos metadados e objetos digitais da coleção, sendo responsáveis também por criar comunidades, gerenciar permissões de acesso, manter o Arca em funcionamento e em segurança física e de dados, e oferecer treinamentos para operação do repositório.

Os bibliotecários gestores dos NAACs têm acesso somente aos registros designados de sua unidade e são responsáveis pela inclusão da produção intelectual da unidade, pela edição, certificação, aprovação e disponibilização dos registros ao público, e pelo gerenciamento da produção sob embargo.

Qualquer pessoa com acesso à internet pode realizar buscas, navegar pelas comunidades e coleções, consultar os registros do Arca, fazer download de documentos e acessar as estatísticas do serviço, sem necessidade de se identificar. Os usuários do serviço também podem se cadastrar no repositório, permitindo-lhes depositar trabalhos, ver depósitos aceitos, editar seu perfil e receber notificações de novos itens adicionados ao Repositório.

3.5 Práticas correntes de preservação

A equipe do Arca participa do Grupo de Trabalho em Preservação Digital do PRESERVO e desenvolve diversas ações de adequação a padrões e requisitos internacionais para preservação digital do acervo, como diagnósticos de confiabilidade do repositório institucional e capacitação da equipe sobre temas ligados à preservação digital.

Entre as ações de preservação atualmente adotadas pelo repositório Arca estão a geração de cópias de segurança (*backup*), a verificação dos arquivos por antivírus e a monitoração da integridade (*checksum*) dos arquivos depositados. A equipe do Arca também iniciou testes para a geração de pacotes de informação para preservação digital a longo prazo, de acordo com as especificações do modelo OAIS.

Backup

A cópia de segurança dos arquivos digitais do Arca é realizada regularmente pela Seção de Infraestrutura da Biblioteca de Manguinhos, no ICICT, sendo uma cópia mensal completa de todos os arquivos do repositório (Backup Full); e cópia diária dos arquivos alterados desde o último backup completo (Backup Incremental). No entanto, o *storage* 3PAR 7200 onde os *backups* do Arca são realizados está atualmente sem contrato de manutenção⁴.

Antivírus

Todos os arquivos depositados no Arca são escaneados pelo antivírus ClamAV (Clam Antivirus), instalado no servidor. O ClamAV é um antivírus gratuito, de código aberto, desenvolvido especialmente para o sistema operacional Linux, e é atualizado diariamente.

Caso seja identificada a presença de vírus em algum arquivo, um alerta é enviado pelo setor de TI para a equipe de coordenadores do Arca (repositorio.arca@fiocruz.br).

Integridade

Para cada objeto digital depositado no repositório, o DSpace gera automaticamente uma soma de verificação MD5 (*checksum*) no momento do depósito e armazena esse valor no sistema. O Arca possui uma rotina que executa a verificação semanal do *checksum* de todos os arquivos depositados. O relatório de resultado da verificação da soma é enviado por e-mail para os coordenadores do Arca (repositorio.arca@fiocruz.br) todos os domingos. Isso garante que o administrador do sistema seja notificado caso alguma soma de verificação indique algum arquivo corrompido ou alterado.

Geração de pacotes de informação

Em dezembro de 2019, o Arca iniciou os testes para extração de pacotes de armazenamento de informação (AIP), utilizando a função *AIP Backup and Restore*⁵ do DSpace, com posterior envio para ingestão pelo software Archivematica, utilizando uma instância de teste da Casa de Oswaldo Cruz (COC), com resultados satisfatórios.

Os AIPs podem ser usados tanto para ingestão por um sistema de preservação digital como para restauração dos arquivos, metadados e estrutura informacional do repositório Arca como um todo ou a nível individual de item, coleção ou comunidade. Os pacotes AIP gerados devem ser armazenados em local geograficamente diverso da instância de produção do Arca.

⁴ Informações do Plano de Backup Arca, disponível em <https://www.arca.fiocruz.br/handle/icict/23055>.

⁵ Mais informações sobre a função do DSpace *AIP Backup & Restore* estão disponíveis em: <https://wiki.lyrasis.org/display/DSDOC6x/AIP+Backup+and+Restore#AIPBackupandRestore-ExportModes&Options>

4. AVALIAÇÃO DE RISCOS

A seguir, estão elencadas as possíveis ameaças identificadas em relação à preservação do acervo digital do Arca. Para cada fator identificado, buscamos avaliar o nível do risco com base na probabilidade de sua ocorrência e nos potenciais impactos ou consequências, bem como os mecanismos de controle existentes para auxiliar na mitigação do risco. Os resultados são apresentados na Tabela 4.

A análise dos riscos foi feita utilizando uma escala de 5 níveis (muito baixo, baixo, médio, alto, muito alto) atribuídos a probabilidade e ao impacto de cada risco. A soma desses níveis determina o nível de risco (Tabela 3).

Essa análise deve nortear futuras decisões do Comitê Gestor do Arca e auxiliar na proposta e implementação de ações para aperfeiçoamento das rotinas e práticas da equipe de operação do Arca.

Tabela 3 - Níveis de risco

Impacto \ Probabilidade	Muito baixo	Baixo	Médio	Alto	Muito alto
Muito baixo	1	2	3	4	5
Baixo	2	4	5	6	7
Médio	3	5	6	7	8
Alto	4	6	7	8	9
Muito alto	5	7	8	9	10

Nível de risco: 1-2 Muito baixo 3-4 Baixo; 5-6 Médio; 7-8 Alto; 9-10 Muito alto

Tabela 4 - Planilha de identificação de riscos

Descrição do risco	Probabilidade	Impacto	Nível de Risco	Controles existentes	Mitigação
Falha no equipamento de armazenamento do backup, que atualmente encontra-se sem contrato de manutenção.	Baixa	Alto Caso haja problemas com o equipamento de armazenamento, não será possível restaurar o backup ou garantir uma janela de retenção, e não há outro equipamento para substituir ou garantir a segurança do backup.	Médio	Existe um backup do Arca nos servidores da sala-cofre da Fiocruz, que pode ser utilizado em caso de catástrofe com as cópias de segurança na Biblioteca de Mangueiras. Parte do conteúdo digital da instância de produção do Arca também é replicado nas instâncias de homologação e/ou treinamento, porém essa não é uma opção segura de backup a ser considerada.	Delinear um plano de ação para mitigar o risco identificado.
Corrupção da cadeia de bits de um objeto digital depositado no Arca	Baixa	Médio O arquivo corrompido pode ser substituído por uma cópia digital do autor ou recuperada a partir de outra fonte.	Médio	Uma soma de verificação (checksum) do objeto é armazenada na ingestão e é realizada uma verificação periódica de integridade de todos os objetos depositados no repositório.	Caso algum arquivo seja corrompido, o administrador poderá recuperar o arquivo não corrompido de uma das cópias de backup anteriores ou solicitar o reenvio do arquivo pelo autor.
Não existe um plano de contingência para restauração do sistema ou dos objetos digitais em caso de desastres	Baixa	Muito alto	Alto	O Arca possui backups na Biblioteca de Mangueiras e nos servidores da sala-cofre da Fiocruz que podem ser acionados para a recuperação do conteúdo do DSpace. No entanto, não existe um plano de ação documentado ou outras opções de recuperação do Arca.	Delinear um plano de ação para mitigar o risco identificado.

5. PLANEJAMENTO DE OBSOLESCÊNCIA

5.1 Requisitos para preservação

Todos os registros depositados no Arca deverão ser contemplados para preservação digital de longo prazo, incluindo tanto os objetos digitais como seus respectivos metadados. A prioridade de preservação, no entanto, é dada às categorias de documentos sob depósito mandatório – teses, dissertações e artigos de periódicos.

Todas as características de apresentação dos documentos devem ser preservadas, bem como garantidas sua integridade e acessibilidade. A estrutura organizacional dos registros (em comunidades e coleções) também deve ser mantida, a fim de permitir o restabelecimento completo ou parcial do repositório Arca, em caso de falha ou perda dos dados.

5.2 Estratégias de preservação avaliadas

A fim de atender aos requisitos para preservação digital, e dentro dos recursos disponíveis para a gestão do repositório Arca, foram avaliadas as seguintes alternativas:

Estratégia 1 - Criação manual de arquivos para geração de SIP

A criação manual de pacotes de informação para preservação foi feita a partir da exportação de metadados e objetos digitais do DSpace.

Foram extraídos os metadados de 5 registros de uma coleção do repositório Arca, posteriormente copiados manualmente para uma planilha em formato CSV pré-formatada com os campos para importação dos metadados por um software de preservação digital, sendo incluídos campos adicionais customizados.

Devido a limitações no processamento do formato CSV, todas as vírgulas presentes nos metadados tiveram que ser substituídas por ponto e vírgula, gerando alterações significativas no conteúdo. Após essa etapa, a planilha CSV e os objetos digitais foram organizados em uma estrutura de diretórios pré-definida.

Realizamos um teste de ingestão dos dados em uma instância de testes do software Archivematica, que resultou na criação com sucesso de um pacote SIP. No entanto, os metadados desses registros contêm erros derivados das alterações feitas na entrada de dados na planilha.

Vantagens: permite a ingestão de dados do Arca em um sistema de preservação digital; mantém a estrutura de arquivos do Arca.

Desvantagens: processo manual, demanda tempo e trabalho excessivos; não mantém a integridade dos metadados; necessita um software de preservação digital ou ferramentas adicionais para verificar a validade dos pacotes criados.

Estratégia 2 - Exportação de pacotes de armazenamento de informação (AIP)

O DSpace permite exportar itens individuais, coleções, comunidades ou todo o conteúdo do repositório em formato de pacotes de armazenamento de informação (AIP) compatíveis com o modelo OAIS, utilizando a função AIP Backup and Restore⁶.

⁶ Ver documentação em <https://wiki.lyrasis.org/display/DSDOC6x/AIP+Backup+and+Restore>.

Como exemplo, para exportar uma coleção em formato AIP, é utilizado o comando na forma:

```
./dspace packager -d -a -t AIP -e <username> -i 12345 aip_colecao.zip
```

, onde "12345" é o handle da coleção e "aip_colecao.zip" é o nome do arquivo AIP resultante.

Para exportar todo o site do DSpace para AIP, incluindo todas as comunidades, subcomunidades, coleções e itens no mesmo diretório do AIP do site, é usado o comando na forma a seguir:

```
./dspace packager -d -a -t AIP -e <username> -i 1234/0 sitewide-aip.zip
```

, onde "1234/0" é o handle do site (seguido de /0) e "sitewide-aip.zip" é o nome do arquivo AIP resultante.

Os arquivos AIP do DSpace podem ser armazenados local ou remotamente para preservação e restauração dos itens (metadados e objetos digitais) e da estrutura das coleções. Segundo a documentação do DSpace, os AIPs exportados também podem ser utilizados como pacotes de submissão (SIP) a um sistema de preservação digital, ou como pacotes de disseminação (DIP). Por padrão, o DSpace valida todo o conteúdo do AIP no momento da exportação, garantindo que todos os AIPs exportados estejam em conformidade com a estrutura de metadados METS e reportando erros, se houver.

Um teste de exportação de AIP foi realizado com uma coleção do Arca composta de 5 registros e gerou um arquivo ZIP contendo os metadados e objetos digitais de cada registro individual estruturados em formato METS e um documento com os metadados em formato METS da coleção da como um todo.

O arquivo ZIP resultante foi transferido para uma instância de testes do software Archivematica, que executou os seguintes processos, culminando na geração de um pacote SIP sem erros, apto para a gestão da preservação digital a longo prazo:

- Aprovação da entrega
- Verificação dos requisitos de conformidade
- Renomeação dos arquivos com atribuição de um identificador exclusivo
- Criação e checagem de somas de verificação (*checksum*)
- Criação dos arquivos XML do METS
- Quarentena de arquivos, quando necessário
- Identificação de formatos de arquivo
- Extração de arquivos compactados
- Verificação de vírus
- Transferência dos arquivos para o diretório de processamento
- Criação do SIP

Com base no resultado dos testes, pudemos comprovar que os arquivos AIP gerados pelo DSpace são uma solução viável:

1. Para armazenamento em servidor externo e uso na restauração do site completo, de coleções e de itens individuais do repositório;
2. Para envio a um sistema de gestão de preservação (como o Archivematica) para execução das demais ações de preservação digital, incluindo a criação e ingestão de SIPs.

Considerando que a instituição não dispõe ainda de um sistema de gestão de preservação em operação, somente a primeira opção se mostra factível no momento. Dentro da estrutura do modelo OAIS, o DSpace funcionará como ferramenta de submissão e acesso aos metadados e objetos digitais, e o armazenamento será realizado em um servidor externo. A etapa de gerenciamento da preservação ocorrerá somente após a implementação de um sistema de preservação.

No entanto, ainda é necessário realizar testes da função *AIP Backup and Restore* na instalação DSpace do repositório Arca para comprovar a viabilidade de restauração dos registros.

Vantagens: Processo rápido, que pode ser automatizado; permite exportar coleções completas e restaurar toda a coleção ou somente itens corrompidos; AIP pode ser mantido em armazenamento local/remoto para ser reimportado pelo DSpace quando necessário ou usado em conjunto com um software de preservação digital; AIP inclui todos os metadados administrativos e estruturais do DSpace; AIP pode ser usado como alternativa ao backup do Arca; conformidade com o padrão OAIS.

Desvantagens: Demanda espaço de armazenamento remoto ou a adoção de um sistema de preservação digital; ausência de testes de reingestão dos pacotes pelo DSpace.

5.3 Seleção da estratégia

A estratégia selecionada consiste na exportação de pacotes de armazenamento de informação para preservação a longo prazo (AIP) pelo DSpace, a serem enviados para armazenamento externo.

A equipe do Arca estuda adotar como solução futura a migração retrospectiva de todo o conteúdo do DSpace para o sistema de preservação digital a ser implantado pela Fiocruz, incluindo a exportação/ingestão automatizada de novos itens depositados no Arca.

5.3.1 Fluxograma de trabalho

A Tabela 5 apresenta de forma detalhada todas as tarefas necessárias para realização das ações de preservação, indicando a periodicidade e o responsável por cada tarefa.

Tabela 5 - Ações, tarefas, periodicidade e responsabilidades do plano de preservação digital do repositório Arca

Ação de preservação	Tarefa	Periodicidade	Responsável
Validação de metadados	Verificar a presença de metadados mínimos em todos os registros ingressados no repositório	Contínuo	- NAACs - Bibliotecários das unidades - Bibliotecário/Bolsista Arca
	Realizar curadoria dos metadados de registros existentes	Contínuo	- Bibliotecário/Bolsista Arca
Padronização de formatos	Verificar se os formatos de arquivo depositados estão em conformidade com as diretrizes do Manual de Digitalização e do Plano Operativo do Arca	Contínuo	- NAACs - Bibliotecários das unidades - Bibliotecário/Bolsista Arca
Nomenclatura de arquivos	Verificar se o nome dos objetos digitais depositados está de acordo com o padrão previsto no Manual de Tratamento de Dados do Arca	Contínuo	- NAACs - Bibliotecários das unidades - Bibliotecário/Bolsista Arca
Verificação de antivírus	Executar programa antivírus em todos os objetos digitais depositados no repositório	Diária	- Equipe TI/ICICT
	Enviar alerta caso seja identificada a presença de vírus em algum arquivo	Quando aplicável	- Equipe TI/ICICT
	Revisar alerta de antivírus e providenciar substituição/remoção do arquivo infectado	Quando aplicável	- Bibliotecário Arca
Monitoramento de integridade	Executar rotina de verificação do checksum de todos os arquivos depositados no repositório	Semanal	- Analista de sistemas Arca
	Enviar alerta caso haja ocorrência de anomalia na verificação de checksum	Quando aplicável	- Bibliotecário Arca
	Providenciar substituição/remoção do objeto digital corrompido	Quando aplicável	- Bibliotecário Arca
Cópias de segurança	Executar rotina de backup conforme especificado no Plano de Backup do Arca	Mensal (completo) Diária (incremental)	- Equipe TI/ICICT
Exportação do AIP	Executar rotina de exportação de AIP do site completo	Semanal	- Analista de sistemas Arca
Arquivamento do AIP	Enviar arquivos AIP para armazenamento externo	Semanal	- Analista de sistemas Arca

6. RECUPERAÇÃO DE DESASTRES

No caso de falha ou perda de todo ou parte do conteúdo digital do acervo do repositório Arca, os metadados e arquivos poderão ser restaurados a partir das cópias de segurança (backup) do DSpace.

Alternativamente, os pacotes de armazenamento de informação (AIP) criados a partir da função *AIP Backup and Restore* do DSpace podem ser usados para a restauração de todo o conteúdo do repositório, bem como de um único item, coleção ou comunidade, incluindo os objetos digitais, os metadados descritivos, estruturais e administrativos, e as configurações de pessoas, grupos e permissões.

7. REVISÃO DO PLANO DE PRESERVAÇÃO DIGITAL

O presente documento deve ser revisto e atualizado pelo Comitê Gestor do Arca em período não superior a dois anos desde a data de publicação da última versão, desde que não haja outros fatores que demandem a sua revisão antecipada.

Exemplos de ocorrências que podem exigir uma atualização extemporânea do Plano são:

- Atualização das políticas institucionais de operação, acesso, preservação e outras relacionadas ao repositório Arca;
- Mudança estrutural nas instâncias de gestão e operação do repositório Arca;
- Mudança ou atualização do sistema de repositório institucional;
- Inclusão ou exclusão de tipologias de documento e formatos de arquivo no sistema;
- Atualização de padrões e melhores práticas de preservação digital;
- Adoção de novas rotinas que impactem no fluxo de trabalho da preservação digital.

8. DOCUMENTOS RELACIONADOS

[Manual de digitalização da Fiocruz](#) (2018)

[Manual de tratamento de dados do Arca](#) (2019)

[Plano operativo: Arca repositório institucional](#) (2017)

[Política de preservação de acervos científicos e culturais da Fiocruz](#) (2018)

Programa de preservação digital da Fiocruz (em preparação, 2020)

