Fluxo de Preservação Dataverse com o Archivematica

Esta página define os requisitos e designs para integração com o Dataverse. A partir do Archivematica v. 1.15.1, a integração com o repositório SciELO Data como um tipo de fonte de transferência oferece suporte à seleção e ao processamento de conjuntos de dados de pesquisa do Dataverse.

Conforme os detalhes nos arquivos de recursos abaixo, a integração foi projetada com o seguinte escopo de uso:

- A integração atual pressupõe um usuário que tenha uma conta com uma instância do Dataverse e tenha gerado uma chave de API associada, e o mesmo usuário (ou um diferente, autorizado) que tenha acesso a uma instância do Archivematica e serviço de armazenamento que esteja conectado a esse Dataverse por meio da chave de API.
- Você pode ler mais na documentação do Dataverse em "Permissões do Dataverse raiz" sobre usuários, administradores e categorias de superusuário que podem impactar o acesso aos conjuntos de dados do Dataverse por meio da API.
- Presume-se que o usuário tenha obtido os direitos necessários para processar e armazenar arquivos de conjuntos de dados no Dataverse para preservação e tenha acesso apropriado ao conjunto de dados e/ou arquivos associados com base nos direitos relacionados à sua chave de API do Dataverse (veja acima).
- Presume-se que o preservador esteja interessado em selecionar conjuntos de dados específicos em um Dataverse para preservação. SIPs e seus AIPs resultantes são criados a partir de versões atuais de conjuntos de dados do Dataverse com um ou mais arquivos associados nesse conjunto de dados. Um conjunto de dados é, portanto, equivalente a um SIP. Arquivos individuais não podem ser selecionados para preservação, nem versões mais antigas de arquivos. No entanto, os usuários podem fazer uso das funções de avaliação do Archivematica para selecionar arquivos individuais em um conjunto de dados específico para criar um AIP final.
- Atualmente, uma função para automatizar a ingestão de todos os conjuntos de dados em um Dataverse não foi desenvolvida.

Apresentando: Preservando um Conjunto de Dados do Dataverse

Rondineli é um usuário do Archivematica
E ele quer preservar um conjunto de dados publicado em um Dataverse

Definições:

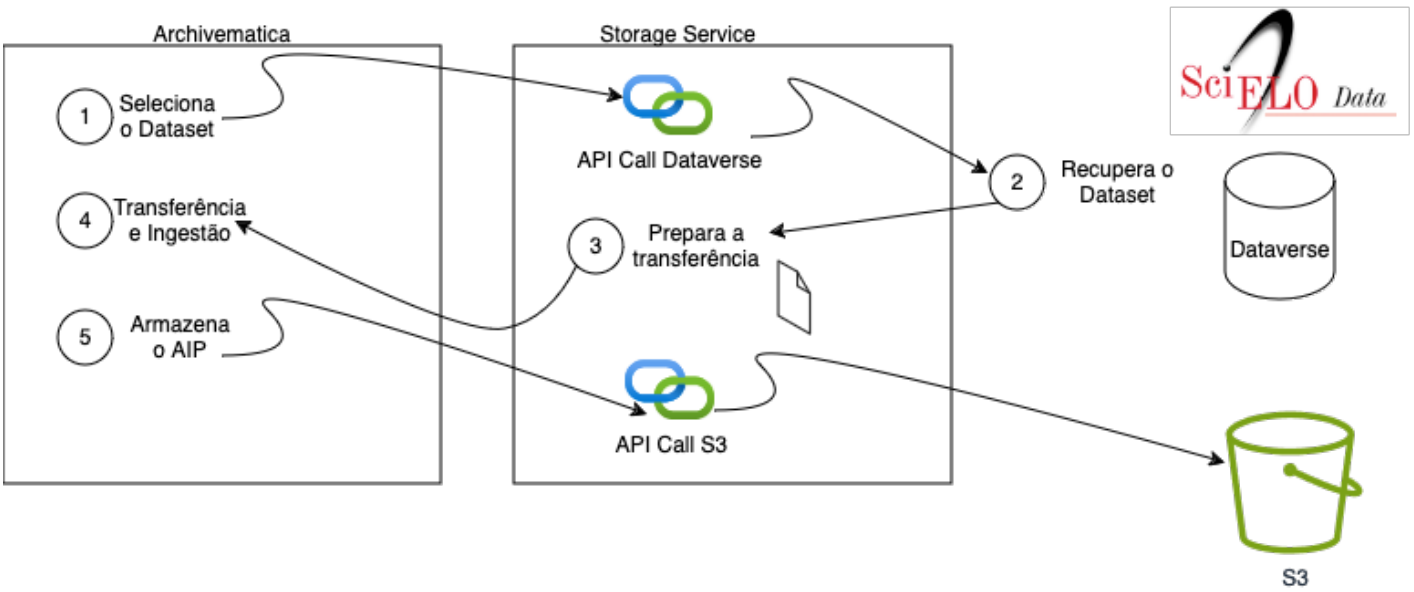
- Conjunto de dados do Dataverse: Um conjunto de dados que foi publicado em um Dataverse, incluindo todos os

arquivos originais enviados para o Dataverse e quaisquer arquivos derivados criados pelo Dataverse.
- METS do Dataverse: Um arquivo de metadados usando o padrão METS que descreve um conjunto de dados;
  incluindo metadados descritivos, lista de todos os objetos no conjunto de dados, sua estrutura
  e relacionamentos entre si.

Cenário: Seleção manual do conjunto de dados

- Dado que o Serviço de armazenamento está configurado para se conectar a um repositório do Dataverse
- E o conjunto de dados foi publicado no Dataverse
- Quando o usuário seleciona o tipo de transferência “Dataverse”
- E o usuário seleciona o conjunto de dados a ser preservado
- E o usuário insere o <Nome da transferência>
- E o usuário insere o <Número de acesso>
- E o usuário clica no botão “Iniciar transferência”
- Então o Archivematica copia os arquivos do Dataverse para um diretório de processamento local
- E o microsserviço Aprovar transferência pede ao usuário para aprovar a transferência
- E o usuário seleciona sim
- E o microsserviço Verificar conformidade de transferência cria o METS do Dataverse
- E os arquivos de metadados do Dataverse são gerados e incluídos em um diretório de metadados
- E o microsserviço Verificar conformidade de transferência confirma que esta é uma transferência válida do Dataverse
- E o microsserviço Verificar somas de verificação de transferência confirma que as somas de verificação fornecidas pelo Dataverse correspondem às geradas para cada arquivo no conjunto de dados
- E o arquivo Mets do AIP inclui os eventos gerados pelo Dataverse
- E o concluído O AIP é armazenado no local de armazenamento especificado do Dataverse

Diagrama do Fluxo de Preservação



1) O usuário seleciona o conjunto de dados Quando o Serviço de armazenamento é configurado para se conectar ao Dataverse, o Navegador de transferência no Painel exibirá uma lista de todos os Locais de origem de transferência do Dataverse. Os locais de origem de transferência podem ser configurados para filtrar termos de pesquisa ou um Dataverse específico. Os usuários podem navegar pelos conjuntos de dados disponíveis, selecionar um e definir o Tipo de transferência como Dataverse.

2) O serviço de armazenamento recupera o conjunto de dados Os serviços de armazenamento usam a API do Dataverse para recuperar o conjunto de dados selecionado. As credenciais da API são armazenadas no Espaço do serviço de armazenamento.

3) Preparar transferência

O Archivematica cria um arquivo de metadados chamado agents.json que inclui as informações do agente configuradas no serviço de armazenamento. Essas informações são usadas para preencher os detalhes do agente PREMIS nos arquivos METS. Consulte Dataverse#agents.json para obter mais detalhes.

Quando um conjunto de dados inclui um "pacote" de arquivos relacionados para dados tabulares, ele é fornecido como um arquivo .zip. O Archivematica extrai todos os arquivos em pacotes neste estágio. Outros arquivos .zip não são afetados e podem ser extraídos ou não usando as opções de configuração de processamento padrão.

4) Transferência e ingestão

O Archivematica realiza processos de transferência e ingestão usando as opções de configuração de processamento padrão. O processamento adicional para conjuntos de dados do Dataverse inclui a criação de um METS do Dataverse que descreve o conjunto de dados conforme fornecido pelo Dataverse verificação de fixidez de arquivos usando somas de verificação fornecidas pelo Dataverse incluindo metadados do Dataverse (do METS do Dataverse) no METS AIP final.

5) Armazene o AIP

O AIP é armazenado em qualquer local que tenha sido configurado. O SciELO Data armazena seus AIPs em um local S3.

Fluxos de trabalho relacionados a pacotes

Pacotes enviados pelo usuário É comum que os usuários do Dataverse façam "dupla compactação" de arquivos ao fazer upload de arquivos para conjuntos de dados. Essa é a prática de compactar arquivos e depois compactá-los novamente uma segunda vez. O Dataverse sempre descompacta os pacotes enviados, mas se os usuários fizerem dupla compactação, eles podem economizar o trabalho de fazer upload de muitos arquivos um por um. Os usuários do Archivematica podem escolher se desejam que esses pacotes sejam extraídos e/ou excluídos posteriormente, definindo a configuração de processamento correspondente apropriada.

Pacotes derivados criados pelo Dataverse Um segundo conjunto de pacotes é criado pelo Dataverse na forma de pacotes derivados. Derivativos são cópias de arquivos em formato tabular que o Dataverse cria a partir de arquivos enviados pelo usuário. O Dataverse entrega esses pacotes ao Archivematica como pacotes zip. Consulte Pacotes para arquivos de dados tabulares para obter mais detalhes abaixo. Consulte o guia do Dataverse sobre ingestão tabular para obter documentação adicional. Esses pacotes são sempre extraídos pelo Archivematica por padrão. Definir a configuração de processamento para não extrair pacotes não funcionará para esse tipo de transferência.

Problemas conhecidos que afetam as transferências

A tabela a seguir resume os problemas conhecidos que afetam o sucesso das transferências individuais:

Problema	Descrição	Etapa de Falha	Mensagem
O conjunto de dados não possui arquivos	Conjuntos de dados que não contêm arquivos (ou seja, apenas metadados) resultarão em uma transferência com falha	Verificar conformidade de transferência: converter estrutura do Dataverse	ConvertDataverseError: Error adding Dataset files to METS

O conjunto de dados possui arquivos com valores de soma de verificação em branco	Uma transferência com falha ocorrerá se o conjunto de dados tiver arquivos com valores de checksum em branco (um problema conhecido para certos tipos de arquivos que foram depositados no Dataverse v3.6 ou anterior). Um usuário pode contornar esse problema selecionando o tipo de transferência "Padrão" e processando a transferência normalmente. No entanto, o arquivo METS não conterá metadados descritivos e os checksums do Dataverse não serão validados. Os administradores podem desejar solucionar problemas de checksums em branco em suas instâncias do Dataverse para corrigir esse problema.	Verificar conformidade de transferência: converter estrutura do Dataverse	ValueError: Must provide both checksum and checksumtype, or neither. Provided values: and MD5
Conjunto de dados que possui arquivos que falharam durante o upload de ingestão tabular do Dataverse	Uma transferência com falha ocorrerá se houver arquivos que falharam no processo de upload de ingestão tabular no Dataverse (por exemplo, resultados em arquivos .RData ausentes e .tab derivados).	Analisar arquivos externos: Analisar XML do Dataverse METS	ParseDataverseError: Exiting. Returning the database objects for our Dataverse files has failed.

O conjunto de dados tem pacotes derivados e “Excluir pacotes após extração” está definido na configuração de processamento	Se o usuário estiver executando o Archivematica versão 1.15.1, e a transferência contiver arquivos derivados (ou seja, arquivos que foram carregados por meio do processo de ingestão tabular no Dataverse) e a opção “excluir pacotes após a extração” for selecionada na configuração de processamento do Archivematica, a transferência falhará. Isso ocorre porque os arquivos .RData contidos como parte dos arquivos derivados são pacotes e serão excluídos. Esse problema foi corrigido no Archivematica 1.9. A solução alternativa se estiver executando o Archivematica 1.8 é selecionar ‘não’ como a opção na configuração de processamento.	Analisar arquivos externos: Analisar XML do Dataverse METS	IntegrityError: (1048, "Column 'eventOutcomeDetailNote' cannot be null")
O usuário tenta processar conjunto de dados com arquivos restritos	Permissões para processar conjuntos de dados por meio do Archivematica correspondem a permissões de função associadas a um Dataverse por meio de um token de API. Portanto, arquivos restritos devem ser processados usando um token de API de administrador ou superusuário para quaisquer conjuntos de dados restritos selecionados para transferência. Caso contrário, o processamento desses conjuntos de dados falhará.	Analisar arquivos externos: Analisar XML do Dataverse METS	ParseDataverseError: Exiting. Returning the database objects for our Dataverse files has failed.

O usuário não seleciona o tipo de transferência “Dataverse”	Ao processar um conjunto de dados do Dataverse, os usuários devem selecionar o tipo de transferência “Dataverse” no menu suspenso ao iniciar a transferência. Se um tipo de transferência “padrão” for selecionado, o conjunto de dados poderá ser processado sem metadados descritivos. Se outro tipo de transferência for selecionado, a transferência falhará. Observação: os Termos de Uso do Conjunto de Dados podem existir para arquivos restritos; nesses casos, espera-se que os Termos de Uso sejam respeitados pela(s) pessoa(s) que processam os arquivos no Archivematica. As informações de licença para conjuntos de dados com arquivos restritos não estão mapeadas atualmente para o METS.	Várias	Várias

Conjunto de Dados do Dataverse

Os conjuntos de dados do Dataverse conforme entregues ao Archivematica contêm o seguinte - Os arquivos originais enviados pelo usuário - Um arquivo de metadados agents.json e dataset.json que descreve os arquivos. - Se o usuário enviou dados tabulares, um conjunto de derivados dos arquivos de tabulação originais em vários formatos, juntamente com arquivos de metadados descrevendo os arquivos tabulares. Consulte a documentação do Dataverse para obter mais informações sobre ingestão tabular.

Arquivo de metadados do conjunto de dados - dataset.json

Este arquivo é fornecido pelo Dataverse. Ele contém citações e outros metadados de nível de estudo, um campo entity_id que é usado para identificar o estudo no Dataverse, informações de versão, uma lista de arquivos de dados com seus próprios valores entity_id e somas de verificação md5 para cada arquivo de dados (original). (Atualmente, ele não fornece somas de verificação para derivados ou arquivos de metadados criados pelo dataverse)

Arquivo de metadados de agentes - agents.json

Este arquivo é criado pelo Archivematica. Ele inclui as informações do Agente que são inseridas no Serviço de Armazenamento ao configurar um Local do Dataverse. Para fazer: adicionar link para documentos finais assim que forem atualizados.

Pacotes para arquivos de dados tabulares

Quando o Dataverse ingere algumas formas de dados tabulares, ele cria derivados do arquivo de dados original e arquivos de metadados adicionais. Todos esses arquivos são fornecidos em um pacote como um pacote compactado, contendo:

- O arquivo original carregado pelo usuário;
- Diferentes formatos derivados (alternativos) do arquivo original (por exemplo, arquivo delimitado por tabulação, arquivo de dados R)
- Metadados variáveis (como um arquivo XML do DDI Codebook);
- Citação do arquivo de dados (atualmente no formato RIS ou EndNote XML);

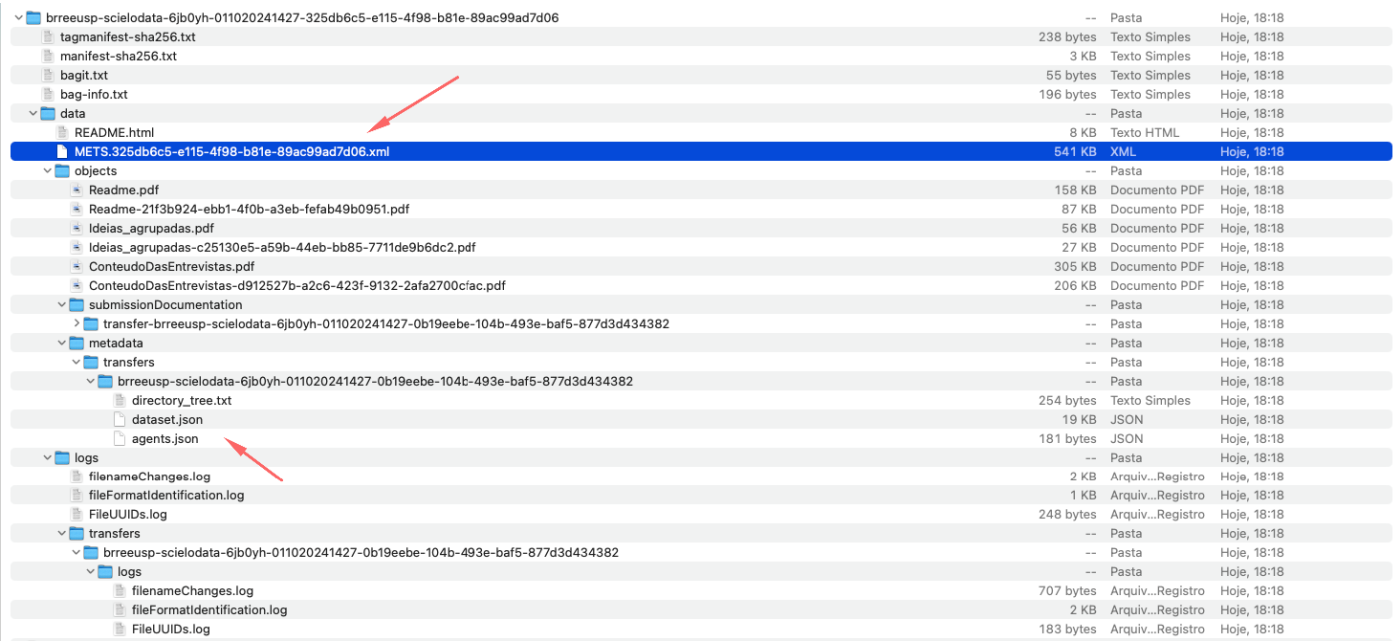
Se o arquivo json tiver content_type de valores separados por tabulação, o Archivematica emite uma chamada de API para download de conteúdo de vários arquivos ("agrupados"). Isso retorna um pacote compactado para arquivos tsv contendo o arquivo .tab, o arquivo original carregado, vários outros formatos derivados, um arquivo DDI XML e citações de arquivo nos formatos Endnote e RIS.

Arquivo METS do Dataverse

O Archivematica gera um arquivo Dataverse METS que descreve o conteúdo do conjunto de dados conforme recuperado do Dataverse. O Dataverse METS inclui:

- metadados descritivos sobre o conjunto de dados, mapeados para o padrão DDI
- uma seção <mets:fileSec> que lista todos os arquivos fornecidos, agrupados por tipo (original, metadados ou derivados)
- uma seção <mets:structMap> que descreve a estrutura dos arquivos conforme fornecidos pelo Dataverse (particularmente útil para entender quais arquivos foram fornecidos em 'pacotes')
- O Dataverse METS é encontrado no AIP final neste local: <Nome do AIP>/data/objects/metadata/transfers/<nome da transferência>/METS.xml (Aqui também é onde você encontrará o arquivo de metadados dataset.json fornecido pelo Dataverse e o arquivo de metadados agents.json criado pelo Archivematica).

Exemplo de arquivo METS do Dataverse



Estudo original do Dataverse recuperado por meio de chamada de API:

- dataset.json (um arquivo JSON gerado pelo Dataverse que consiste em metadados de nível de estudo e informações sobre arquivos de dados)
- Arquivos .pdf (arquivos de dados não tabular)

Arquivo METS Dataverse resultante

- O fileSec no arquivo METS consiste em três grupos de arquivos, USE="original" (os arquivos PDF e SAV); USE="derivative" (os arquivos TAB e R); e USE="metadata" (o arquivo JSON e os três arquivos de metadados do pacote compactado).
- Todos os arquivos descompactados do pacote Dataverse têm um atributo GROUPID para indicar o relacionamento entre eles. Se a transferência tivesse consistido em mais de um pacote, cada conjunto de arquivos descompactados teria seu próprio GROUPID.
- dmdSecs foram gerados:
  - dmdSec_1, consistindo em um pequeno número de termos DDI de nível de estudo

- No structMap e dmdSec_1 são vinculados ao estudo como um todo,

Transferir arquivo METS

Durante o processamento de transferência, um arquivo Transfer METS é criado. Ele é encontrado no AIP final neste local: <Nome do AIP>/data/objects/submissionDocumentation/<nome da transferência>/METS.xml

Arquivo AIP METS

Estrutura básica do arquivo METS

O arquivo Archival Information Package (AIP) METS seguirá a estrutura básica para um arquivo AIP METS padrão do Archivematica descrito em METS. Um novo fileGrp USE="derivative" será

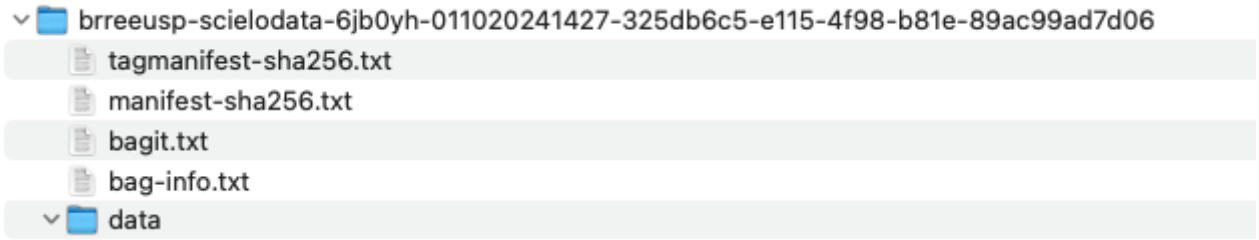
adicionado para indicar TAB, RData e outros derivados gerados pelo Dataverse para arquivos de formato de dados tabulares carregados.

Estrutura AIP

Um Archival Information Package derivado de uma ingestão do Dataverse terá a mesma estrutura básica de um Archivematica AIP genérico, descrito em AIP_structure. Há arquivos de metadados adicionais que são incluídos em um AIP derivado do Dataverse, e cada pacote compactado que é incluído na ingestão resultará em um diretório separado no AIP. A seguir está uma estrutura de amostra.

Estrutura Bag

O Archival Information Package (AIP) é empacotado no formato BagIt da Biblioteca do Congresso e pode ser armazenado compactado ou descompactado:



Estrutura do AIP

Todo o conteúdo do AIP reside no diretório de dados:

```
├── data
│   ├── logs [arquivos de log gerados durante o processamento]
│   │   ├── fileFormatIdentification.log
│   │   ├── filenameChanges.log
│   │   ├── FileUUIDs.log
│   │   └── transfers
│   │       └── transfer-brreeusp-scielodata-6jb0yh-011020241427-0b19eebe-104b-493e-baf5-
877d3d434382
│   │           └── logs
│   │               ├── filenameChanges.log
│   │               ├── fileFormatIdentification.log
│   │               └── FileUUIDs.log
│   ├── METS.325db6c5-e115-4f98-b81e-89ac99ad7d06.xml [o arquivo AIP METS]
│   ├── objects [um diretório contendo os objetos digitais que estão sendo preservados, além
de seus metadados]
│       ├── Readme.pdf [an original file from Dataverse]
│       ├── Readme-21f3b924-ebb1-4f0b-a3eb-fefab49b0951.pdf [an original file from Dataverse]
```

```
|       ├── metadata
|       |   └── transfers
|       |       └── brreeusp-scielodata-6jb0yh-011020241427-0b19eebe-104b-493e-baf5-
877d3d434382
|       |           ├── agents.json [see Dataverse#agents.json]
|       |           ├── dataset.json [see Dataverse#dataverse.json]
|       |           └── directory_tree.txt [see ]
|       └── submissionDocumentation
|           └── transfer-brreeusp-scielodata-6jb0yh-011020241427-0b19eebe-104b-493e-baf5-
877d3d434382
|               └── METS.xml [o arquivo padrão Transfer METS descrito acima]
```

Estrutura do arquivo AIP METS

O arquivo AIP METS registra informações sobre o conteúdo do AIP e indica os relacionamentos entre os vários arquivos no AIP. Um arquivo AIP METS de exemplo seria estruturado da seguinte forma:

```
METS dmdSec [seção de metadados descritivos]
-link to dataset.json
METS dmdSec [seção de metadados descritivos]
-link to DDI.XML arquivo criado para arquivo derivado como parte do pacote
METS amdSec [seção de metadados administrativos, uma para cada arquivo original, derivado e
normalizado no AIP]
-techMD [metadados técnicos]
--PREMIS metadados técnicos sobre um objeto digital, incluindo informações sobre o formato do
arquivo e metadados extraídos
-digiprovMD [metadados de proveniência digital]
--PREMIS event: derivation (para formatos derivados)
-digiprovMD [metadados de proveniência digital]
--PREMIS event:ingestion
-digiprovMD [metadados de proveniência digital]
--PREMIS event: unpacking (para arquivos agrupados)
-digiprovMD [metadados de proveniência digital]
--PREMIS event: cálculo do resumo da mensagem
-digiprovMD [metadados de proveniência digital]
--PREMIS event: verificação de vírus
-digiprovMD [metadados de proveniência digital]
--PREMIS event: identificação de formato
-digiprovMD [metadados de proveniência digital]
--PREMIS event: verificação de fixidez (se o arquivo vier do Dataverse com uma soma de
```

```
verificação)
-digiprovMD [metadados de proveniência digital]
--PREMIS event: normalização (se o arquivo for normalizado para um formato de preservação durante o processamento do Archivematica)
-digiprovMD [metadados de proveniência digital]
--PREMIS event: criação (se o arquivo for um mestre de preservação normalizado gerado durante o processamento do Archivematica)
-digiprovMD
--PREMIS agent: organização
-digiprovMD
--PREMIS agent: software
-digiprovMD
--PREMIS agent: Usuário do Archivematica
METS fileSec [seção de arquivo]
-fileGrp USE="original" [grupo de arquivos]
--arquivos originais enviados para o Dataverse
-fileGrp USE="derivative"
--arquivos tabulares derivados gerados pelo Dataverse
-fileGrp USE="submissionDocumentation"
--METS.XML (arquivo METS de transferência padrão do Archivematica listando o conteúdo da transferência)
-fileGrp USE="preservation"
--mestres de preservação normalizados gerados durante o processamento do Archivematica
-fileGrp USE="metadata"
--dataset.json
--DDI.XML
--xcitation-endnote.xml
--xcitation-ris.ris
METS structMap [mapa estrutural]
-estrutura de diretório do conteúdo do AIP
```

Referência

https://wiki.archivematica.org/Dataverse#Setting_up_the_Integration

Revision #5
Created 9 October 2024 18:32:14 by Rondineli G. Saad
Updated 6 November 2024 17:07:37 by Rondineli G. Saad