HIPÁTIA
MODELO DE PRESERVAÇÃO

ibict

Estudo do Bagit

PRESIDÊNCIA DA REPÚBLICA

Jair Messias Bolsonaro
Presidente da República

Hamilton Mourão
Vice-Presidente da República

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÕES
Marcos Cesar Pontes
Ministro da Ciência, Tecnologia e Inovações.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA
Cecília Leite Oliveira
Diretora

Reginaldo de Araújo Silva
Coordenação de Administração – COADM

Gustavo Saldanha
Coordenação de Ensino e Pesquisa, Ciência e Tecnologia da Informação – COEPPE

José Luis dos Santos Nascimento
Coordenação de Planejamento, Acompanhamento e Avaliação – COPAV

Anderson Itaborahy
Coordenador-Geral de Pesquisa e Desenvolvimento de Novos Produtos - CGNP

Bianca Amaro de Melo

Coordenadora-Geral de Pesquisa e Manutenção de Produtos Consolidados - CGPC

Tiago Emmanuel Nunes Braga
Coordenador-Geral de Tecnologias de Informação e Informática – CGTI

Alexandre Faria de Oliveira
Coordenador de Governança em Tecnologias para Informação e Comunicação - COTIC

EQUIPE TÉCNICA
Diretora do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict)
Cecília Leite Oliveira

Coordenador-Geral de Tecnologias de Informação e Informática (CGTI)
Tiago Emmanuel Nunes Braga

Coordenador do Projeto
Alexandre Faria de Oliveira

Autores
Tatiana Canelhas Pinataro
Daniel Monteiro
Alexandre Faria de Oliveira

Normalização
Marilete da Silva Pereira

Projeto gráfico e capa
Alisson Eugênio da Costa

Este Relatório de Técnico é um produto do projeto Preservação digital e gestão arquivística apoiada no aprimoramento da implantação do modelo RDCArq.

Ref. TJDFT - Processo SEI no 01302.000067/2021-45 e 01302.000389/2020-11
Ref. FUNDEP 28209

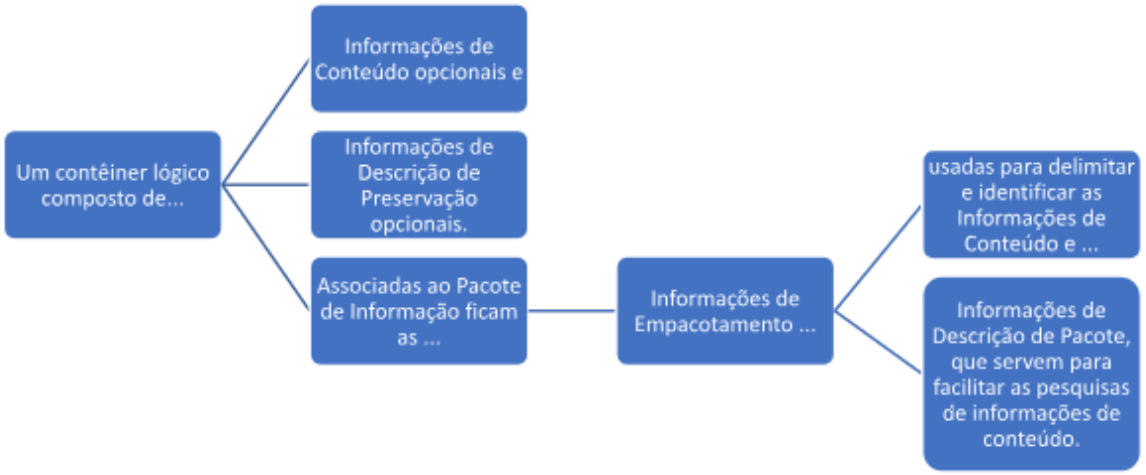
Setor de Autarquias Sul Quadra 05 Lote 06, Bloco H – 5o andar
Cep:70.070-912 – Brasília, DF
Telefones: 55 (61) 3217-6360/55 /(61)3217-6350
www.ibict.br

Sumário

1 PLANO DE PRESERVAÇÃO

O *Open Archival Information Systems* – OAIS, ISO 14721:2012, é o modelo referencial para recomendações práticas de um repositório digital confiável, que "descreve as funções de um repositório digital e os metadados necessários para a preservação e o acesso dos materiais digitais gerenciados pelo repositório, que constituem um modelo funcional e um modelo de informação" (CONARQ, RDC-Arq) e estabelece que a forma mais segura de troca de informações entre sistemas é através de pacotes de informação que envolvem os objetos digitais e seus metadados.

De acordo com o *REFERENCE MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS), RECOMMENDED PRACTICE CCSDS 650.0-M-2* (MAGENTA BOOK, 2012), o com conceito de Pacote de Informação é:



Fonte: Magenta Book, pg. 12 (tradução nossa)

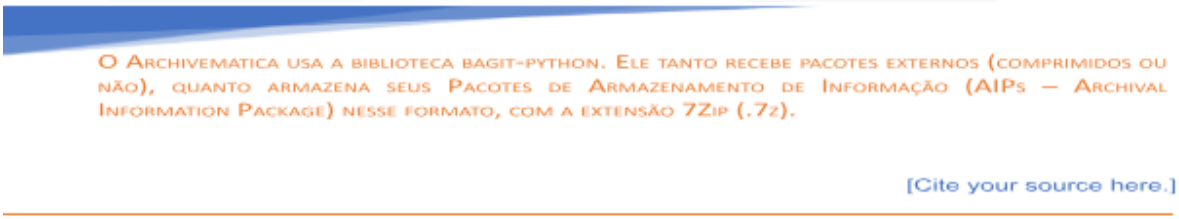
Existem três tipos de pacotes de informações no modelo de referência OAIS:

- SIP – Pacote de Informação para Submissão (*Submission Information Package*) – admissão do pacote ao repositório digital confiável contendo a informação do produtor ao arquivo.
- AIP – Pacote de Informação para Arquivamento (*Archival Information Package*) – arquivamento do pacote ao repositório digital confiável.
- DIP – Pacote de Informação para Disseminação (*Dissemination Information Package*) – acesso do conteúdo do pacote ao sistema de acesso ao ser requisitado por um usuário.

O Archivematica é o sistema que será instalado no Arquivo Nacional para ser usado em seu ambiente de preservação de documentos digitais. Trata-se de um software livre desenvolvido pela empresa canadense Artefactual, com interface web, e baseado no modelo OAIS. Para recolhimento dos processos do Sistema Eletrônico de Informações (SEI) do Arquivo Nacional ao seu arquivo digital, o IBICT usará o *BagIt* como formato de empacotamento para seus objetos digitais e metadados, por ser este o padrão adotado pelo archivematica.

2 BAGLT

BagIt é o formato padronizado de empacotamento utilizado pelo Archivematica, criado pela Biblioteca do Congresso Americano sob licença pública. Ele organiza de forma hierárquica os documentos digitais junto aos seus metadados.



2.1 Pacotes de Transferência de Informação

O archivematica recebe pacotes de outros sistemas, compactados (*zipped bags*) ou descompactados (*unzipped bags*), mas o barramento foi adaptado para trabalhar apenas com pacotes compactados. Esses pacotes foram nomeados para Pacotes de Transferência de Informação (*Transfer Information Package – TIP*). Os TIPs devem cumprir alguns requisitos para que não gerem erros no archivematica, interrompendo assim sua transferência[1]:

- É proibido adicionar arquivos invisíveis antes da criação do *bag*;
- Os requisitos de codificação relatados no arquivo *bagit.txt* não podem ser diferentes da codificação dos caracteres encontrados no arquivo *bag-info.txt*;
- É proibido editar qualquer arquivo após a criação do *bag*;
- Deve-se tomar cuidado com a estrutura de pastas do *bag*.

[1]ttps://www.archivematica.org/en/docs/archivematica-1.11/user-manual/transfer/bags/#bag-structure-requirements

No barramento, o *bag* é construído utilizando a ferramenta *BagIt-Python*[2], que está de acordo com as especificações de um *BagIt*. Os formatos podem ser .zip, .tgz ou tar.gz.

2.1.1 Grupo de Formatos

O BagIt compactado a ser enviado para o archivematica com os dados extraídos do SEI será a estruturado conforme exemplo abaixo:

2.1.1.2 bag-info.txt

Arquivo texto que conterá informações que serão indexadas no *ElasticSearch*, tornando-se assim pesquisáveis no Archivematica após arquivamento do AIP. "Os campos no arquivo *bag-info.txt* são serializados como XML no campo sourceMD do METS e vinculados ao diretório de objetos do AIP."[3] (*Manual Archivematica*).

Todos os metadados são opcionais e podem ser repetidos (com exceções de alguns metadados reservados). Um metadado deve consistir em seu nome, seguido de dois pontos ":", depois um único caractere de espaço em branco (espaço ou

[2] BagIt-Python é uma biblioteca Python e comandos para trabalhar com pacotes no estilo BagIt. (https://github.com/LibraryOfCongress/bagit-python)

[3] https://www.archivematica.org/en/docs/archivematica-1.11/user-manual/transfer/bags/#index-and-search-bag-metadata

tabulação) e um valor que é terminado com um LF (l*ine feed - Linux*), um CR (*carriage return - MAC*) ou um CRLF (Windows).

> "A etiqueta NÃO DEVE conter dois pontos (:), LF ou CR. O rótulo deve conter caracteres de espaço em branco lineares, mas NÃO PODEM começar ou terminar com espaço em branco. É RECOMENDADO que as linhas não excedam 79 caracteres de comprimento. Os valores longos PODEM ser continuados na próxima linha inserindo um LF, CR ou CRLF, e então recuando a próxima linha com um ou mais caracteres de espaço em branco (espaços ou tabulações)."[4]

Existe uma lista de metadados reservados importantes (ainda que opcionais), como:

- Source-Organization: instituição que está transferindo os dados.
- Organization-Address: endereço da instituição.
- Contact-Name: nome da pessoa responsável na instituição pelo conteúdo transferido.
- Contact-Phone: número de telefone (no formato internacional) da pessoa responsável na instituição pelo conteúdo transferido.
- Contact-Email: e-mail da pessoa responsável na instituição pelo conteúdo transferido
- External-Description: Uma breve explicação do conteúdo e proveniência.
- Bagging-Date: Data (AAAA-MM-DD) que o conteúdo foi transferido.
- External-Identifier: Um identificador fornecido pelo remetente para o bagIt.
- Bag-Size: O tamanho do bagIt transferido (pode ser tamanho aproximado), seguido pela abreviatura do dimensionamento (MB – megabytes, GB – gigabytes ou TB – terabytes).
- Payload-Oxum: A "soma octetstream" da bag. É destinada a detectar bagIts incompletas antes de realizar a validação da soma de verificação.
- Bag-Group-Identifier: um identificador único fornecido pelo remetente para o grupo de bagIts aos quais logicamente pertence.

[4] https://tools.ietf.org/html/rfc8493

- Bag-Count: Dois números separados pela palavra "de", "N de T", onde T é o número total de bagIts em um grupo de bags e N é o número ordinal dentro do grupo.
    - Se este metadados está presente, é recomendado incluir também o metadado *Bag-Group-Identifier*.
- Internal-Sender-Identifier: um identificador específico do remetente alternativo para o conteúdo e / ou bagIt.
- Internal-Sender-Description: uma explicação do remetente do conteúdo e proveniência.

Segue abaixo um exemplo de como o bag-info.txt pode ser escrito:

```
Source-Organization: Arquivo Nacional
Organization-Address: Rio de Janeiro
Contact-Name: Djalma Brito
Contact-Phone: +55 21 3408-555-1212
Contact-Email: ej@an.gov.br
External-Description: Imagens TIFF em escala de cinza não compactadas da coleção
de Yoshimuri ...
Bagging-Date: 2020-01-15
External-Identifier: spengler_yoshimuri_001
Bag-Size: 260 GB
Payload-Oxum: 279164409832.1198
Bag-Group-Identifier: spengler_yoshimuri
Bag-Count: 1 of 15
Internal-Sender-Identifier: /storage/imagens/yoshimuri
Internal-Sender-Description: TIFFs em escala de cinza não compactados criados a
partir de microfilmes e são ...
```

Quando preservada no arquivo METS XML do AIP resultante, as informações acima são representadas da seguinte forma:

```
<mets:amdSec ID="amdSec_14">
  <mets:sourceMD ID="sourceMD_1">
    <mets:mdWrap MDTYPE="OTHER" OTHERMDTYPE="BagIt">
      <mets:xmlData>
        <transfer_metadata>
          <Source-Organization>Arquivo Nacional</Source-Organization>
          <Organization-Address>Rio de Janeiro</Organization-Address>
          <Contact-Name>Djalma Brito</Contact-Name>
          <Contact-Phone>+55 21 3408-555-1212</Contact-Phone>
          <Contact-Email> ej@an.gov.br </Contact-Email>
          <External-Description> Imagens TIFF em escala de cinza não compactadas
da coleção de Yoshimuri ...</External-Description>
          <Bagging-Date>2020-01-15</Bagging-Date>
          <External-Identifier>spengler_yoshimuri_001</External-Identifier>
          <Bag-Size>260 GB</Bag-Size>
          <Payload-Oxum>279164409832.1198</Payload-Oxum>
          <Bag-Group-Identifier>spengler_yoshimuri</Bag-Group-Identifier>
          <Bag-Count>1 of 15</Bag-Count>

<Internal-Sender-Identifier>/storage/images/yoshimuri</Internal-Sender-Identifier
>
          <Internal-Sender-Description>TIFFs em escala de cinza não compactados
criados a partir de microfilmes e são ...</Internal-Sender-Description>
        </transfer_metadata>
      </mets:xmlData>
    </mets:mdWrap>
  </mets:sourceMD>
</mets:amdSec>
```

2.1.1.2 bagit.txt

O arquivo bagit.txt deve conter exatamente duas linhas, na seguinte ordem:

- BagIt-Vertion: X.Y
- Tag-File-Character-Encoding: <ENCODING>

Onde “X” é a maior e “Y” a menor versão de BagIt e encoding é a codificação de caracteres utilizada. Podemos citar como exemplo de bagit.txt a imagem abaixo:

```
BagIt-Version: 1.0
Tag-File-Character-Encoding: UTF-8
```

2.1.1.3 manifest-<algorithm>.txt

Trata-se de um arquivo de texto que listam os arquivos que estão dentro da pasta /data e suas hashs (*checksums*) correspondentes que foram geradas usando um algoritmo de soma de verificação criptográfica específico. Cada linha do arquivo de carga útil tem o seguinte formato:

```
CHECKSUM FILENAME
```

Onde FILENAME é o caminho de um arquivo relativo ao diretório base e CHECKSUM é uma soma de verificação (*hash*) codificada de acordo com o algoritmo usado no arquivo. Apenas o caractere de barra ('/') pode ser usado como

separador de caminho em FILENAME. Um ou mais caracteres de espaço em branco (espaços ou tabulações) separam CHECKSUM de FILENAME.

Obs.: *Manifests* não contabilizam diretórios vazios, pois incluem apenas os nomes de caminho dos arquivos. “Para citar um diretório vazio, um BagIt deve incluir pelo menos um arquivo nesse diretório. Para isso, basta incluir um arquivo de tamanho zero denominado "*.keep*"[5].

Os checksums são verificados durante o microsserviço “*Verify transfer checksums*” na guia *Transfer*. Os arquivos de checksums são colocados no diretório de metadados.

Segue abaixo um exemplo de manifest-md5.txt:

```
49afbd86a1ca9f34b677a3f09655eae9 data/27613-h/images/q172.png
408ad21d50cef31da4df6d9ed81b01a7 data/27613-h/images/q172.txt
```

2.1.1.3 tagmanifest-<algorithm>.txt

Tem o mesmo conceito que o *manifest-<algotithm>.txt*, mas faz o cálculo das hashs que estão fora da pasta /data.

Segue abaixo um exemplo de tagmanifest-md5.txt:

```
27afbd86a1ca9f34b677a3f09655eaa4 manifest-md5.txt
59ad21d50cef31da4df6d9ed81b01b01 bagit.txt
```

2.1.1.4 /data

[5] https://tools.ietf.org/html/rfc8493

Neste diretório, teremos os objetos digitais que serão capturados do ambiente de gestão SEI, que será subdividido entre duas pastas: /access e /objects.

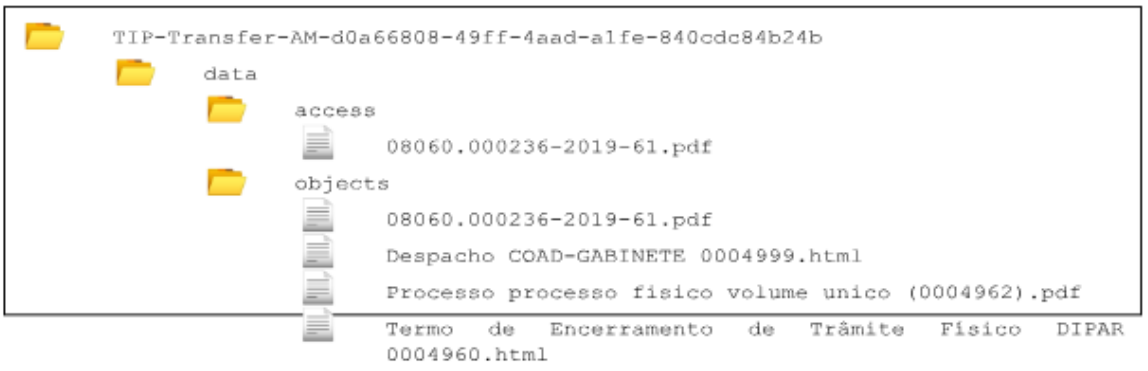
2.1.1.4.1 /data/access

Uma das estratégias do Archivematica para preservar e dar acesso a longo prazo aos arquivos inseridos em seu ambiente é normalizá-los de acordo com as regras *Format Policy Registry* (FPR)[6], já inseridas em seu Plano de Preservação. No entanto, faremos uma normalização[7] prévia do processo do SEI para o arquivo de acesso, que será um único PDF de seus documentos compilados. "O Archivematica pode reconhecer o trabalho de normalização manual e usar as cópias de preservação e acesso em vez de criar novas derivadas"[8]. Para que funcione essa "normalização manual" e que seja entendida no archivematica que as derivadas de acesso já existem, é necessário que se crie um diretório chamado /access dentro do diretório /data com o(s) objeto(s) normalizado(s) dentro. Para que o archivematica localize o arquivo de acesso, o nome dos arquivos devem permanecer os mesmos (arquivo original e arquivo de acesso), conforme imagem abaixo:

[6] O Archivematica gerencia políticas de formato local e externamente por meio de um Registro de Política de Formato (Format Policy Registry – FPR). Essa política indica as ações, ferramentas e configurações a serem aplicadas a um arquivo de um formato específico.

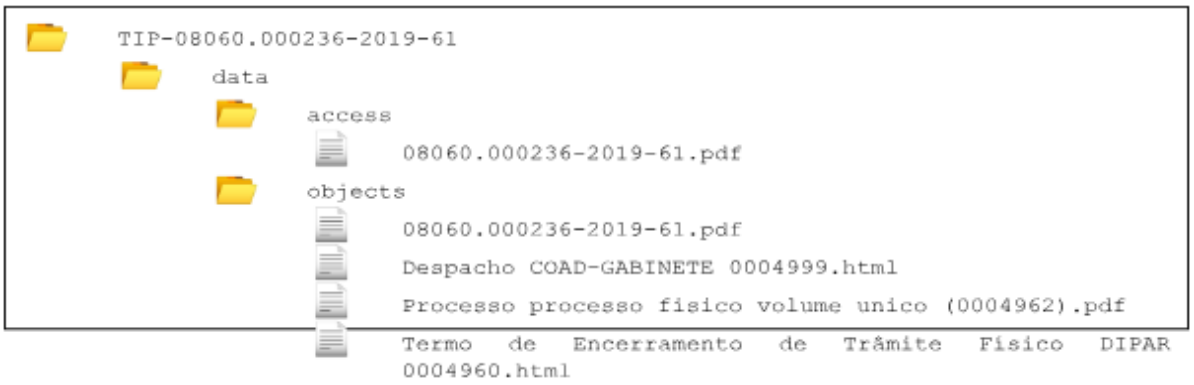
[7] A principal estratégia de preservação do Archivematica é normalizar arquivos após a ingestão. Normalização é o processo de migrar um arquivo de um determinado formato em outro, para uma finalidade declarada (como preservação ou acesso). As cópias de preservação são adicionadas ao AIP (Archival Information Package) e as cópias de acesso são usadas para gerar um DIP (Dissemination Information Package) para ser enviada para o sistema de acesso (por exemplo, o programa Access to Memory, AtoM). Os arquivos originais são sempre mantidos para permitir diferentes ações de preservação no futuro, como normalização para diferentes formatos de arquivo ou emulação (Obs.: O Archivematica mantém o formato original de todos os arquivos ingeridos para suportar estratégias de preservação de migração e emulação).

[8] https://www.archivematica.org/en/docs/archivematica-1.11/user-manual/transfer/transfer/#transfer-derivatives



2.1.1.4.2 /data/objects

Neste diretório serão inseridos todos os objetos digitais que fazem parte do processo que serão capturados do SEI, além do processo com os documentos unificados.

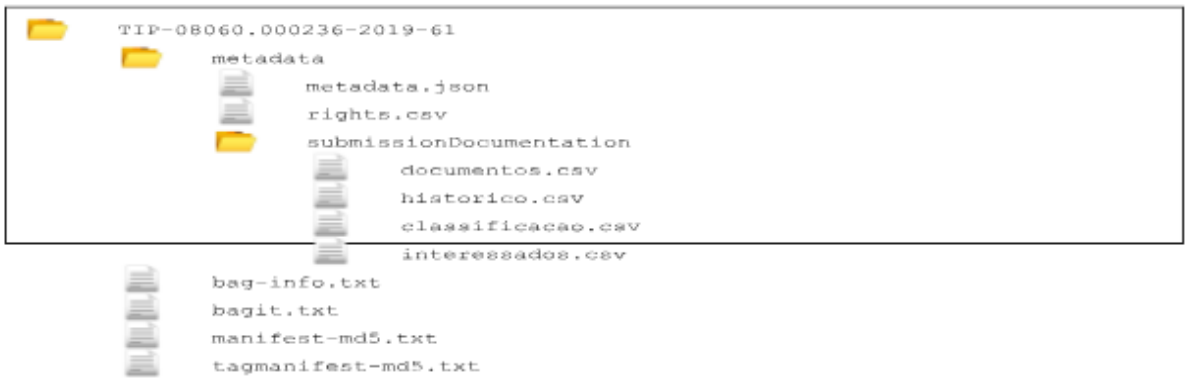


2.1.1.4 /metadata

Neste diretório serão inseridos os metadados de gestão e descrição extraídos do SEI, do processo capturado, e salvos no formato JSON[9]. É importante que o

[9] JSON (JavaScript Object Notation - Notação de Objetos JavaScript) é uma formatação leve de troca de dados. Fonte: https://www.json.org/json-pt.html

nome do arquivo seja metadata.json para que o archivematica o reconheça como o arquivo de metadados e o insira em sua trilha de auditoria AIP METS.



2.1.1.4.2 /metadata/submissionDocumentation

A documentação de submissão (*submission documentation*) é um conceito no Archivematica que contabiliza materiais relacionados aos objetos digitais que estão sendo preservados, mas que não fazem parte estritamente da coleção – por exemplo, acordos de doadores, correspondência sobre materiais, relatórios de conservação etc. Se o Archivematica visualizar que uma transferência inclui documentação de submissão, pode incluir descrições deste material no arquivo AIP METS. Neste diretório serão inseridos os metadados de gestão e descrição também extraídos do SEI, mas referentes a(à):

- Documentos: dados dos documentos que fazem parte do processo;
- Histórico: histórico completo do processo dentro do SEI;
- Classificação: classificações dadas ao processo e aos documentos que fazem parte do processo;
- Interessados: dados das pessoas que fazem parte do quadro de interessados do processo e dos documentos que fazem parte do processo.

2.2 Ingestão do Bag

Essa parte do documento foi baseada no documento sobre a transferência do BagIt para o archivematica criado pela Artefactual.[10]

2.2.1 Requisitos

- Todas as verificações padrões do Bagit são executadas:
    - verifyvalid;
    - checkpayloadoxum;
    - verifycomplete;
    - verifypayloadmanifests;
    - verifytagmanifests.
- O Archivematica diferencia entre os elementos de bag obrigatórios e opcionais de forma que, se os elementos opcionais não estiverem presentes, a execução não falhe no microsserviço de verificação.
- As verificações do BagIt geram arquivos de log que serão adicionados ao diretório de logs da transferência.
- Nenhum novo evento PREMIS é necessário. As verificações BagIt são registradas como verificação de fixidez no PREMIS.

[10] https://wiki.archivematica.org/Bag_ingest

2.2.2 Fluxograma