



Générateur de SIP - Mode d'emploi

Version du 24/04/2017

Valeurs Immatérielles Transférées aux Archives pour Mémoire

Organisation de la présentation

1/ Présentation du générateur

2/ Fonctionnement simple du générateur

3/ Fonctionnement avancé du générateur

Annexe/ Rappels sur le SEDA 2.0.

Présentation du générateur

Objectifs de l'outil

- Faciliter la génération des jeux de tests pour les membres de l'équipe Vitam, les ministères porteurs et les partenaires
- Générer automatiquement, à partir d'une arborescence de fichiers des SIP, pouvant être pris en charge par la solution logicielle Vitam :
 - Compatible NF Z44-022 et standard SEDA v2.0 (conforme schéma .xsd du SEDA 2.0.)
 - Conforme au document de spécification des SIP propre à la solution logicielle Vitam
 - Sans nécessairement avoir besoin d'utiliser un éditeur xml pour créer le bordereau
 - Permettant de générer rapidement un SIP avec des milliers d'unités d'archives et des milliers de fichiers
 - De manière fiable (ne pas copier à la main l'empreinte)

Alimentation du bordereau (1)

- Pour les fichiers
 - Calcul de l'empreinte (avec l'algorithme de hachage paramétrable, par défaut SHA-512) et écriture dans le bordereau
 - Calcul de la taille du fichier et écriture dans le bordereau
 - Récupération dans le bordereau du nom d'origine du fichier et de sa date de dernière modification (FileInfo)
 - Définition d'un usage par défaut pour les fichiers (original numérique = BinaryMaster)
 - Identification du format du fichier en utilisant l'outil Siegfried
 - Rassemblement dans un même groupe d'objets de plusieurs fichiers constituent plusieurs représentations d'une même unité archivistique (ex. une version de conservation et une version de diffusion)
 - *Voir la partie fonctionnement avancée du générateur pour plus de détails*

Alimentation du bordereau (2)

- Pour l'arborescence d'unités de description
 - Création d'une arborescence d'unités archivistiques reprenant l'arborescence du système de fichiers
 - Création des liens entre unités archivistiques et fichiers numériques
 - Transformation des raccourcis vers des répertoires et des fichiers en liens vers ces répertoires et fichiers **NOUVEAU**
 - Gestion des unités archivistiques complexes, avec fichiers rattachés et également arborescence (ex. message électronique) **NOUVEAU**
 - *Voir la partie fonctionnement avancée du générateur pour plus de détails*
 - Indication de niveaux de description par défaut : RecordGrp pour les répertoires, Item pour les fichiers
 - Alimentation automatique du bordereau avec les informations suivantes récupérées de l'arborescence de fichiers
 - *Dates (date de modification des fichiers, dates extrêmes des répertoires)*
 - *Titre (répertoire, fichier)*

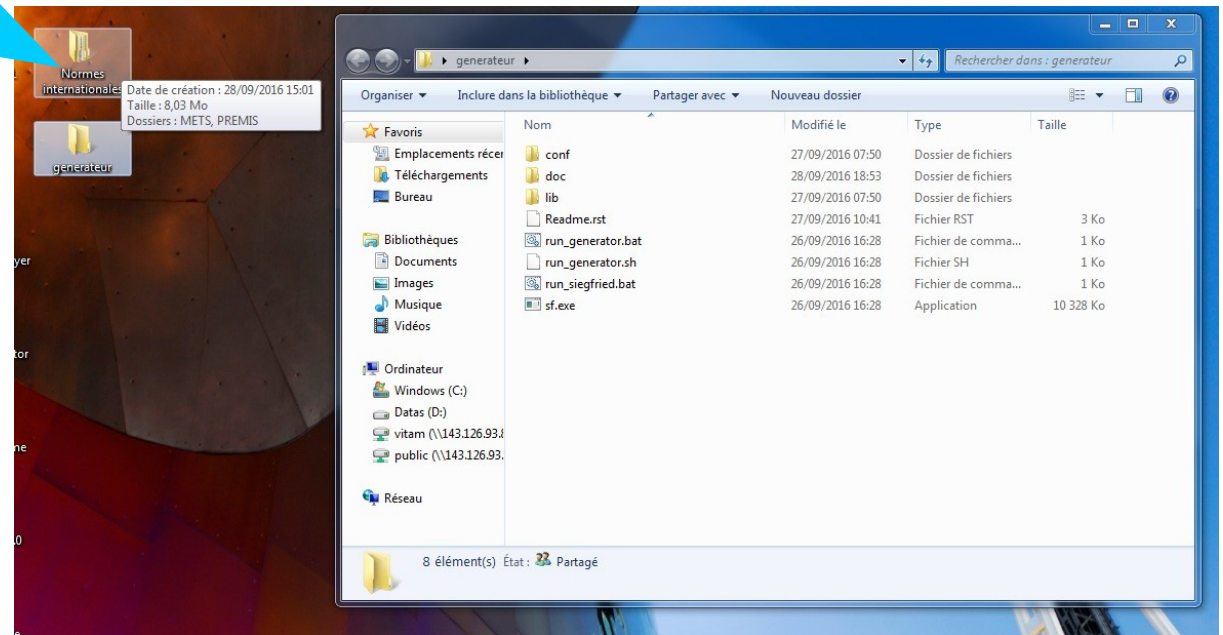
Paramétrage de l'alimentation du SIP

- **Paramétrage de l'en-tête du bordereau**
 - Alimentation de certains champs : Identifiant du SIP, commentaire, contrat d'entrée, identifiant du service effectuant le transfert, identifiant du service d'archives, listes de codes, identifiant du service producteur, identifiant du service versant
 - Définition des fichiers que l'utilisateur ne souhaite pas mettre dans le SIP : par exemple Thumbs.db, *.vcf
- **Paramétrage des métadonnées des répertoires**
 - Alimentation possible de tous les champs existants dans l'ontologie définie dans le SEDA 2.0.
- **Autres paramétrages possibles :**
 - Rejet des fichiers non reconnus par l'outil d'identification des formats Siegfried. Par défaut, ces fichiers sont inclus dans le SIP
 - Rejet des fichiers ayant une extension contenant un caractère « URL-encoded » de type « + ». Par défaut, ces fichiers sont inclus dans le SIP
 - Choix d'un autre algorithme de calcul d'empreinte des fichiers que l'algorithme par défaut (SHA-512)

Fonctionnement simple du générateur

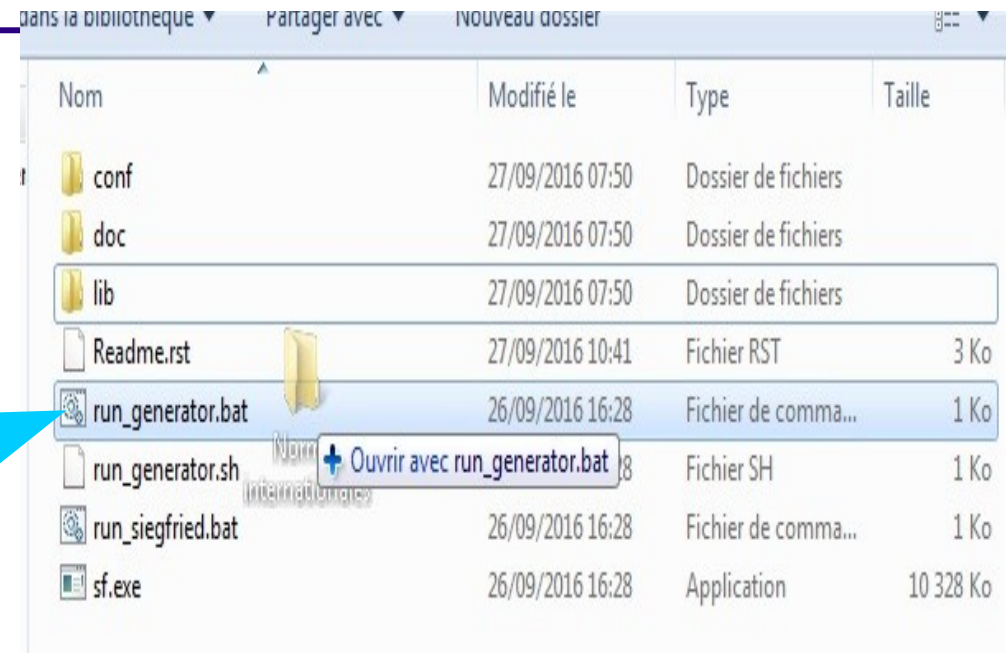
Étape 1 - Identifier le dossier à archiver

- Choisir dans le navigateur la racine de l'arborescence à intégrer dans le SIP
- Attention : si un seul fichier doit être archivé, il convient de l'enregistrer dans un répertoire

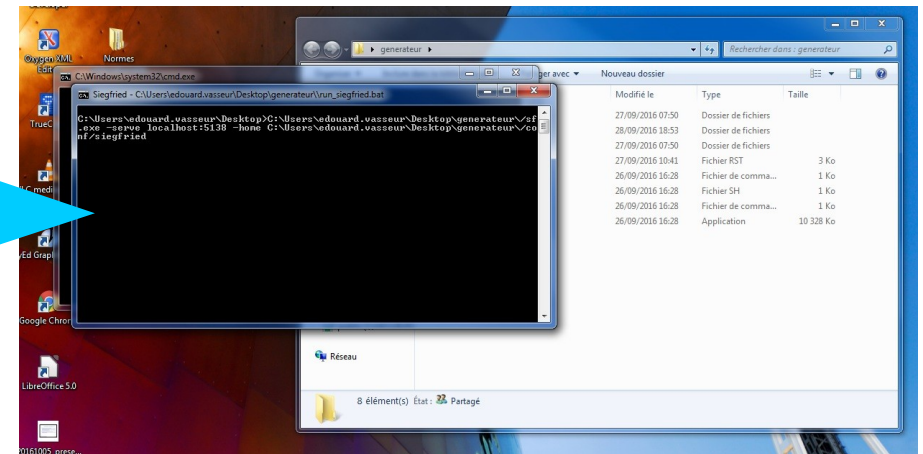


Étape 2 - Lancer l'opération de création du SIP

- Ouvrir le répertoire contenant le générateur
- Faire glisser le dossier sélectionné vers le fichier « run_generator.bat »

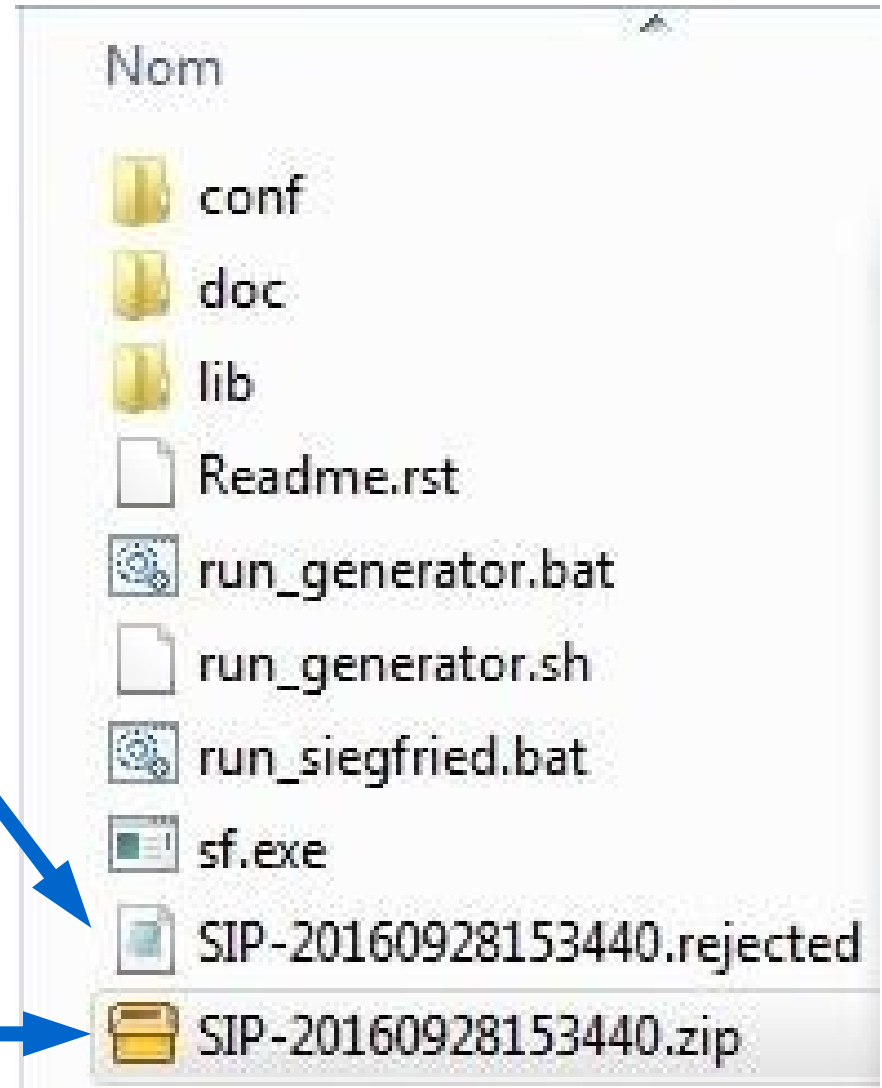


- Deux consoles s'ouvrent. La 2^e console trace les erreurs survenues pendant l'opération. Elle se ferme à la fin de l'opération en tapant sur la touche « entrée »



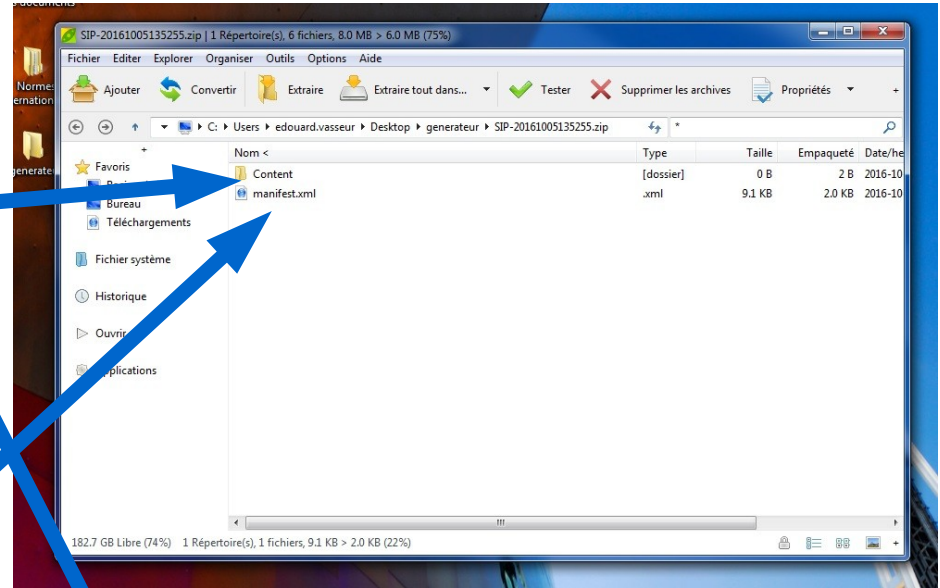
Étape 3 - où retrouver le SIP généré ?

- 2 fichiers SIP ont été créés dans le répertoire contenant le générateur:
 - Le 1^{er} « rejected » relève toutes les erreurs survenues lors de la création du SIP
 - *S'il n'y a pas d'erreur la taille du fichier est 0 Ko*
 - Le 2^e est le SIP à proprement parler avec une extension .zip



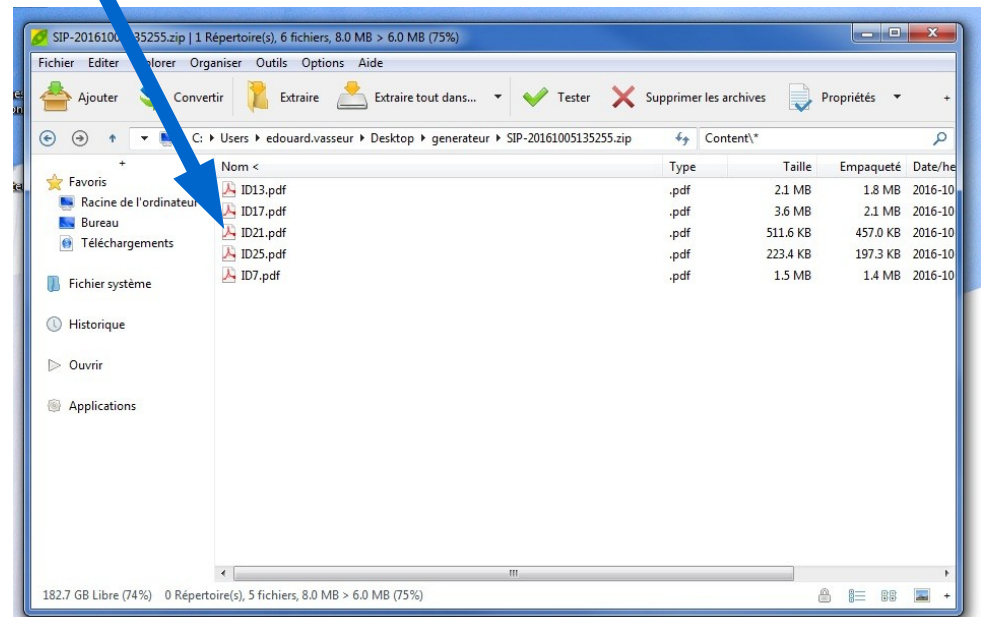
Étape 4 - Consulter le SIP

- Ouvrir le zip et constater qu'il contient bien :
 - Un répertoire **Content** contenant l'ensemble des fichiers de l'arborescence, renommés et à plat
 - Un fichier **manifest.xml** qui correspond au bordereau



NB : le bordereau peut être extrait pour enrichissement (ex. avec un éditeur xml) et réinjecté dans le SIP ensuite (en supprimant la version d'origine)

Le SIP est prêt !



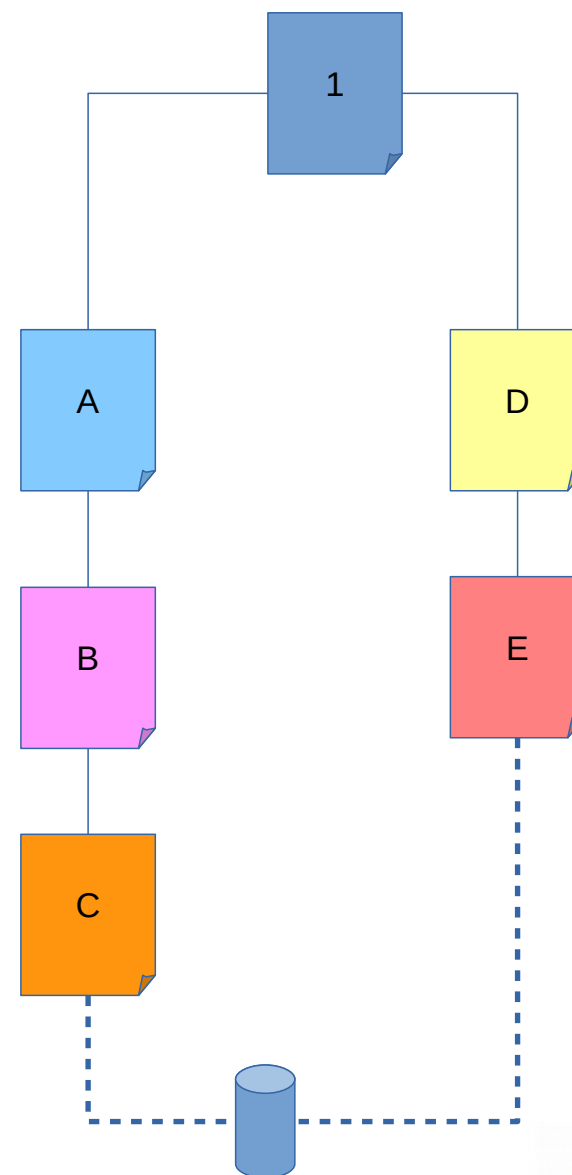
Fonctionnement avancé du générateur

Constituer des jeux de tests « avancés »

Organiser l'arborescence en fonction du comportement attendu

Ex. : Plusieurs branches de l'arborescence pointent vers un même fichier

- Créer 1 répertoire comprenant 2 sous-répertoires
- Prendre le 1^{er} sous-répertoire et y mettre 1 seul sous-sous-répertoire, dans lequel on crée un sous-sous-sous-répertoire dans lequel on positionne le ou les fichiers
- Prendre le 2^e sous-répertoire et y mettre 1 seul sous-sous-répertoire, dans lequel on positionne le ou les mêmes fichiers que dans le sous-répertoire n° 1



Paramétrage de l'en-tête du bordereau (1)

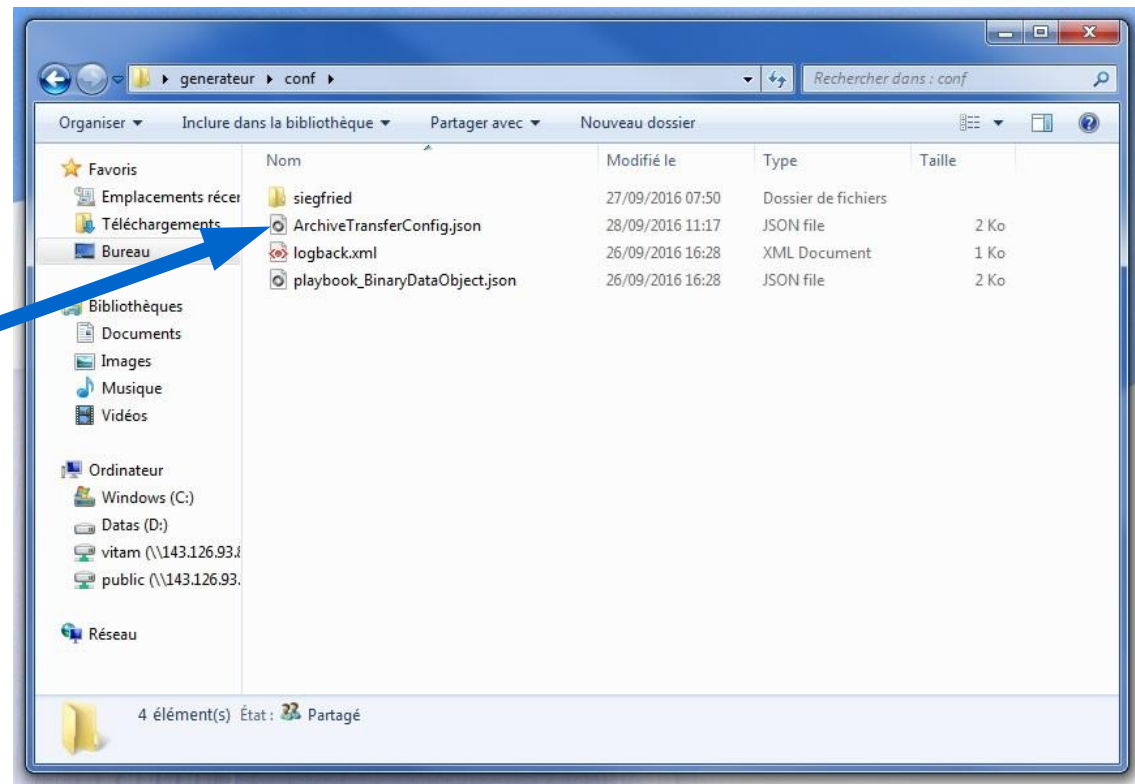
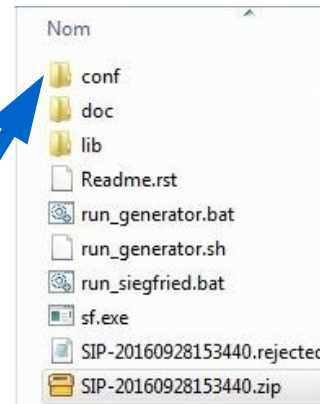
Il est possible de paramétrer l'en-tête du bordereau

Ouvrir le répertoire contenant le générateur

Puis le répertoire Conf

Et utiliser le fichier ArchiveTransferConfig.json

- Copier ce fichier dans le répertoire racine de l'arborescence
- Ouvrir le fichier .json avec un éditeur xml ou un éditeur de texte comme Notepad++



Paramétrage de l'en-tête du bordereau (2)

- Modifier les valeurs des champs qui vous intéressent en respectant la ponctuation (guillemets) :
 - Comment : chaîne de caractère décrivant le transfert
 - MessageIdentifier : identifiant du transfert
 - ArchivalAgreement : contrat d'entrée utilisé
 - CodeListVersions : listes de codes utilisés dans le bordereau
 - ArchivalAgency : identifiant du service d'archives destiné à recevoir le transfert
 - TransferringAgency : identifiant de l'opérateur de transfert à l'origine du transfert
 - OriginatingAgencyIdentifier : identifiant du service producteur
 - SubmissionAgencyIdentifier : identifiant du service versant
- Indiquer les catégories de fichiers que l'utilisateur souhaite exclure du SIP (champ ignore_patterns)
- Sauvegarder les modifications

Paramétrage de l'en-tête du bordereau (3)

```
{
  "Comment" : "2eme SIP",
  "MessageIdentifier" : "MessageIdentifier0",
  "ArchivalAgreement" : "ArchivalAgreement0",
  "CodeListVersions" : {
    "ReplyCodeListVersion" : "ReplyCodeListVersion0",
    "MessageDigestAlgorithmCodeListVersion" : "MessageDigestAlgorithmCodeListVersion0",
    "MimeTypeCodeListVersion" : "MimeTypeCodeListVersion0",
    "EncodingCodeListVersion" : "EncodingCodeListVersion0",
    "FileFormatCodeListVersion" : "FileFormatCodeListVersion0",
    "CompressionAlgorithmCodeListVersion" : "CompressionAlgorithmCodeListVersion0",
    "DataObjectVersionCodeListVersion" : "DataObjectVersionCodeListVersion0",
    "StorageRuleCodeListVersion" : "StorageRuleCodeListVersion0",
    "AppraisalRuleCodeListVersion" : "AppraisalRuleCodeListVersion0",
    "AccessRuleCodeListVersion" : "AccessRuleCodeListVersion0",
    "DisseminationRuleCodeListVersion" : "DisseminationRuleCodeListVersion0",
    "ReuseRuleCodeListVersion" : "ReuseRuleCodeListVersion0",
    "ClassificationRuleCodeListVersion" : "ClassificationRuleCodeListVersion0",
    "AuthorizationReasonCodeListVersion" : "AuthorizationReasonCodeListVersion0",
    "RelationshipCodeListVersion" : "RelationshipCodeListVersion0"
  },
  "ArchivalAgency" : {
    "Identifier" : "Identifier4"
  },
  "TransferringAgency" : {
    "Identifier" : "Identifier5"
  },
  "ManagementMetadata.OriginatingAgencyIdentifier" : "Service_producteur",
  "ManagementMetadata.SubmissionAgencyIdentifier" : "Service_versant",
  "ignore_patterns" : ["Thumbs.db","pagefile.sys"]
}
```

Paramétrage des métadonnées d'un répertoire (1)

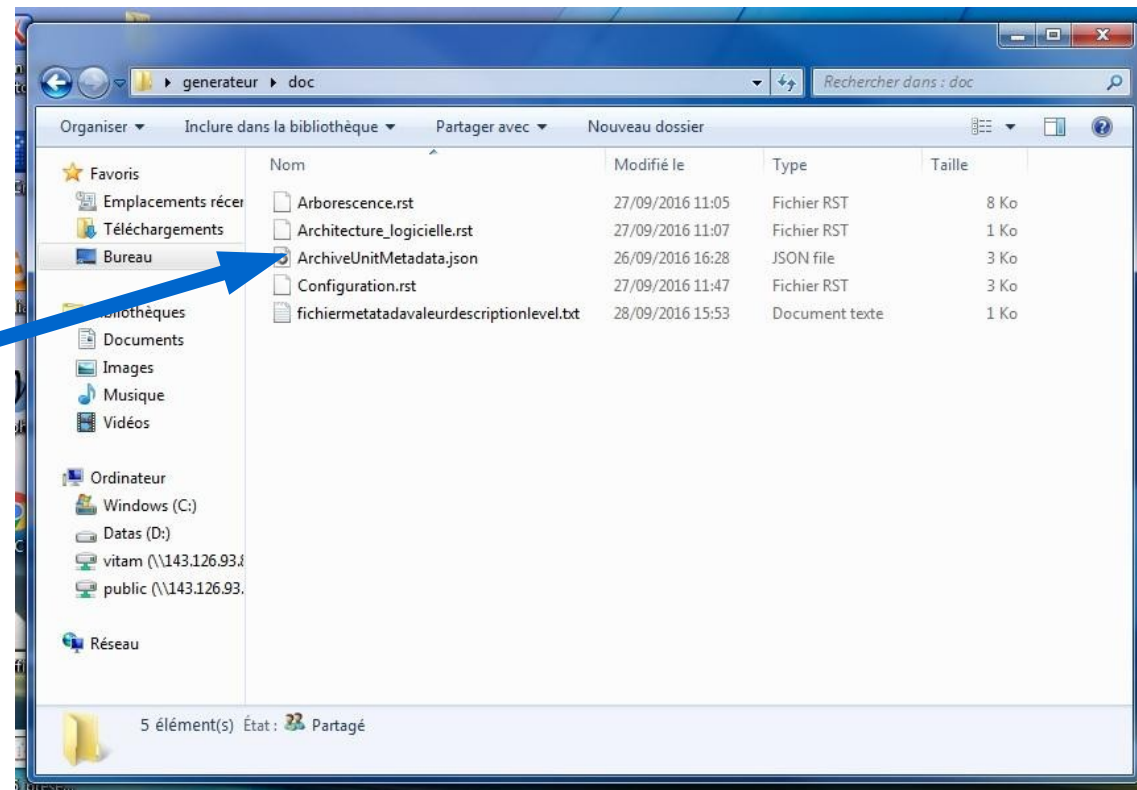
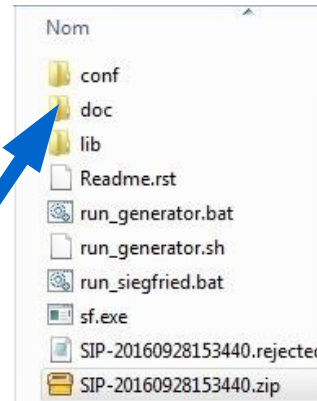
Il est possible de paramétrer l'indexation du répertoire

Ouvrir le répertoire contenant le générateur

Puis le répertoire doc

Et utiliser le fichier ArchiveUnitMetadata.json

- Copier ce fichier dans le répertoire que vous voulez indexer
- Ouvrir le fichier .json avec un éditeur xml ou un éditeur de texte comme Notepad++



Paramétrage des métadonnées d'un répertoire (2)

- Identifier les champs que vous voulez enrichir parmi la liste de tous les champs disponibles dans l'ontologie fournie par le SEDA
- Modifier les valeurs des champs qui vous intéressent en respectant la ponctuation (guillemets)
 - Voir slides correspondantes pour la liste des champs
 - Les champs DescriptionLevel et Title sont obligatoires
- Supprimer les champs que vous ne souhaitez pas voir apparaître dans l'indexation du répertoire
 - Voir slide correspondante pour les points d'attention
- Sauvegarder les modifications

Paramétrage des métadonnées d'un répertoire (3)

```
veUnitMetadata.json | ArchiveTransferConfig.json | ArchiveUnitMetadata.json
{
  "Content" : {
    "DescriptionLevel" : "RECORD_GRP",
    "Title" : [ {
      "Value" : "Titre francais",
      "Lang" : "fr"
    }, {
      "Value" : "English title",
      "Lang" : "en"
    } ],
    "FilePlanPosition" : "Valeur de filePlanPosition",
    "SystemId" : "Valeur de SystemID)",
    "OriginatingSystemId" : "Valeur de OriginatingSystemId",
    "ArchivalAgencyArchiveUnitIdentifier" : "Valeur de archivalAgencyArchiveUnitIdentifier",
    "OriginatingAgencyArchiveUnitIdentifier" : "Valeur de originatingAgencyArchiveUnitIdentifier",
    "TransferringAgencyArchiveUnitIdentifier" : "Valeur de transferringAgencyArchiveUnitIdentifier",
    "Description" : [ {
      "Value" : "Description francaise",
      "Lang" : "fr"
    }, {
      "Value" : "English Description",
      "Lang" : "en"
    } ],
    "Type" : {
      "Value" : "Valeur du type",
      "Lang" : "fr"
    },
    "DocumentType" : {
      "Value" : "fr"
    },
    "Language" : "FR",
    "DescriptionLanguage" : "FR",
    "Status" : "Valeur de Status",
    "Version" : "Valeur de version",
    "Tag" : [ "XML Tag 1 (de type xml:token)", "XML Tag 2 (de type xml:token)" ],
    "Coverage" : {
      "Spatial" : [ {
        "Value" : "Valeur de Spatial",
        "Lang" : "fr"
      } ]
    }
  }
}
```

Paramétrage des métadonnées d'un répertoire (4)

- Champs modifiables
 - DescriptionLevel : niveau de description
 - *Voir slide correspondante pour la liste des valeurs*
 - Title : titre
 - FilePlanPosition : position dans le plan de classement
 - SystemId : identifiant fourni par le SAE
 - OriginatingSystemId : identifiant fourni par le système de production
 - ArchivalAgencyArchiveUnitIdentifier : identifiant fourni par le service d'archives (ex. cote)
 - OriginatingAgencyArchiveUnitIdentifier : identifiant fourni par le service producteur
 - TransferringAgencyArchiveUnitIdentifier : identifiant fourni par le service versant
 - Description : présentation détaillée du contenu
 - Type : type d'information au sens de l'OAIS

Paramétrage des métadonnées d'un répertoire (5)

- Champs modifiables

- DocumentType : type de document
- Language : langue des archives
- DescriptionLanguage : langue des descriptions
- Status : statut du document
- Version : version du document
- Tag : indexation
- Coverage : couverture (Spatial : géographique ; Temporal : chronologique ; jurisdictional : administrative)
- OriginatingAgency : identifiant du service producteur
- SubmissionAgency : identifiant du service versant
- Writer : rédacteur du document
- Source : pour les documents numérisés, référence au document source

Paramétrage des métadonnées d'un répertoire (6)

- Champs modifiables
 - Event : description d'un événement survenu pendant le cycle de vie du document
 - Gps : coordonnées géographiques définies par l'utilisateur

Paramétrage des métadonnées d'un répertoire (7)

- Valeurs possibles pour le champ Niveau de description
 - Fonds : mettre FONDS
 - Sous-fonds : mettre SUBFONDS
 - Classe : mettre CLASS
 - Collection : mettre COLLECTION
 - Série : mettre SERIES
 - Sous-série : mettre SUBSERIES
 - Groupe d'archives : mettre RECORD_GRP
 - Sous-groupe d'archives : mettre SUB_GRP
 - Dossier : mettre FILE
 - Pièce : mettre ITEM

Paramétrage des métadonnées d'un répertoire (8)

- Points d'attention sur la suppression des champs
 - Toujours mettre une virgule à la fin de chaque champ, sauf à la fin du dernier champ
 - les champs Title et Description peuvent avoir plusieurs valeurs. Il faut vérifier qu'il y a bien le bon nombre de caractères spéciaux utilisés pour délimiter le champ au moment de la suppression

Paramétrage des métadonnées d'un répertoire (9)

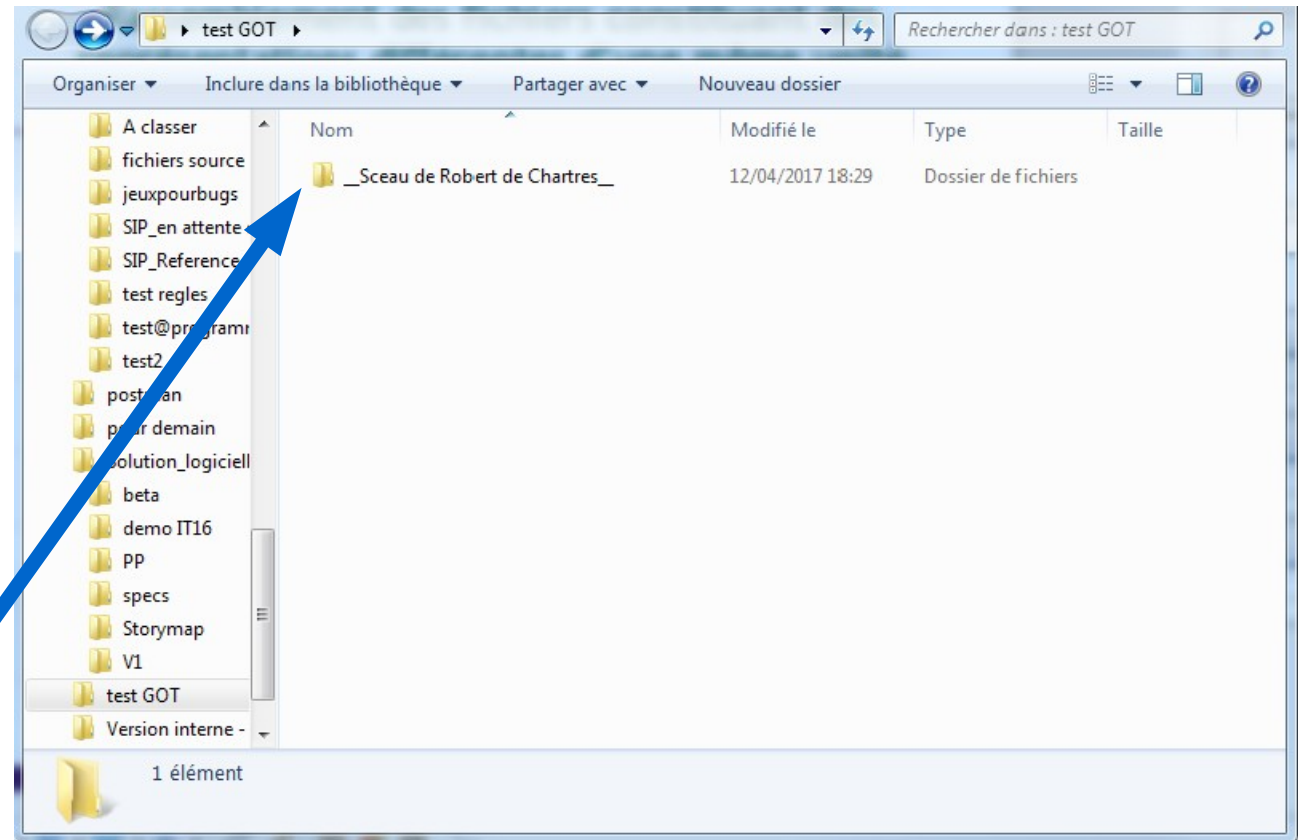
- Points d'attention sur la suppression des champs - exemple

```
{ "Content" : {  
  "DescriptionLevel" : "FILE",  
  "Title" : [ {  
    "Value" : "Documentation sur le standard METS  
  } ],  
  "FilePlanPosition" : "3.1.1.",  
  "OriginatingSystemId" : "24561",  
  "Description" : [ {  
    "Value" : "Documentation récupérée sur le site de la Bibliothèque du Congrès"  
  } ],  
  "Language" : "EN",  
  "OriginatingAgency" : {  
    "Identifier" : {  
      "Value" : "CodeVitam"  
    }  
  },  
  "StartDate" : "2016-09-26T15:34:08.284+02:00",  
  "EndDate" : "2016-09-26T15:34:08.284+02:00"  
}
```

Rassemblement des fichiers constituant des représentations différentes d'une même unité archivistique (1)

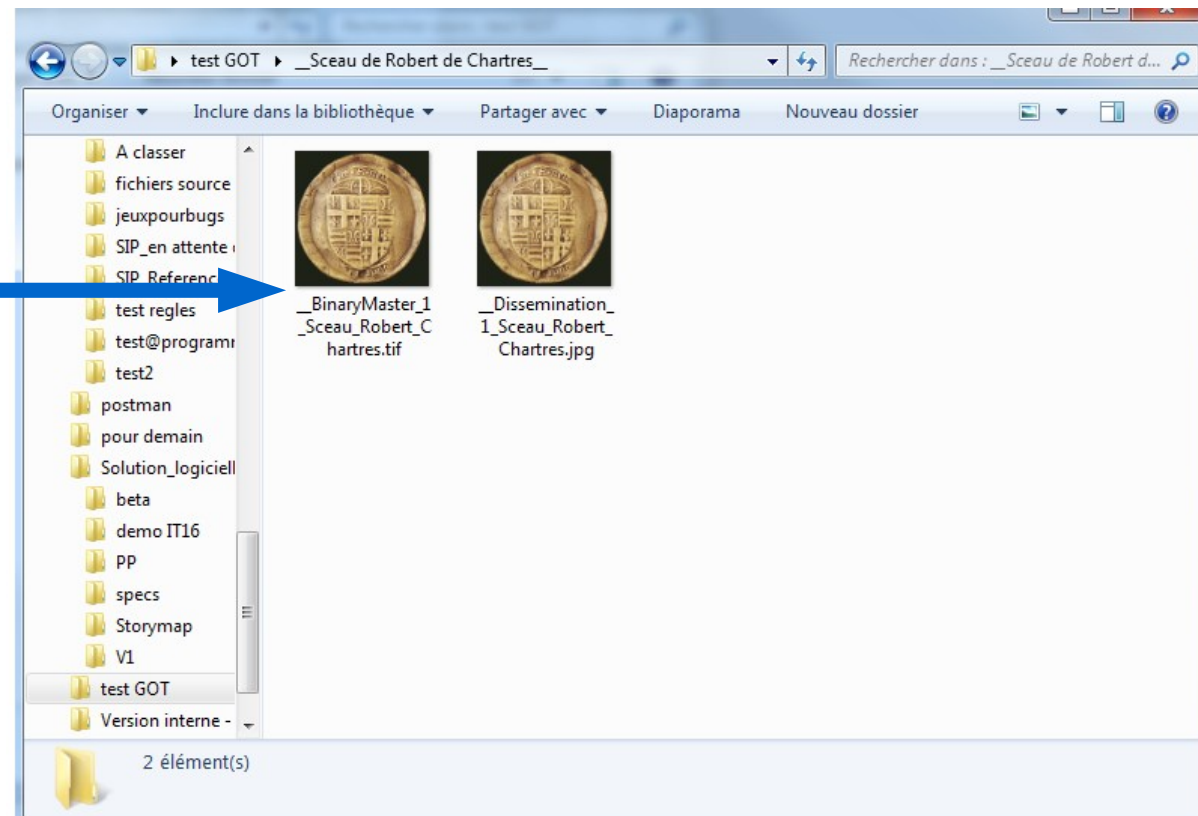
Exemple d'une photographie présente dans 2 formats :

- un format .tiff de conservation
- un format .jpeg de diffusion
- Étape 1 : enregistrer les 2 fichiers dans un même répertoire
- Étape 2 : renommer ce répertoire en ajoutant comme préfixe 2 Underscores et en ajoutant comme suffixe 2 Underscores



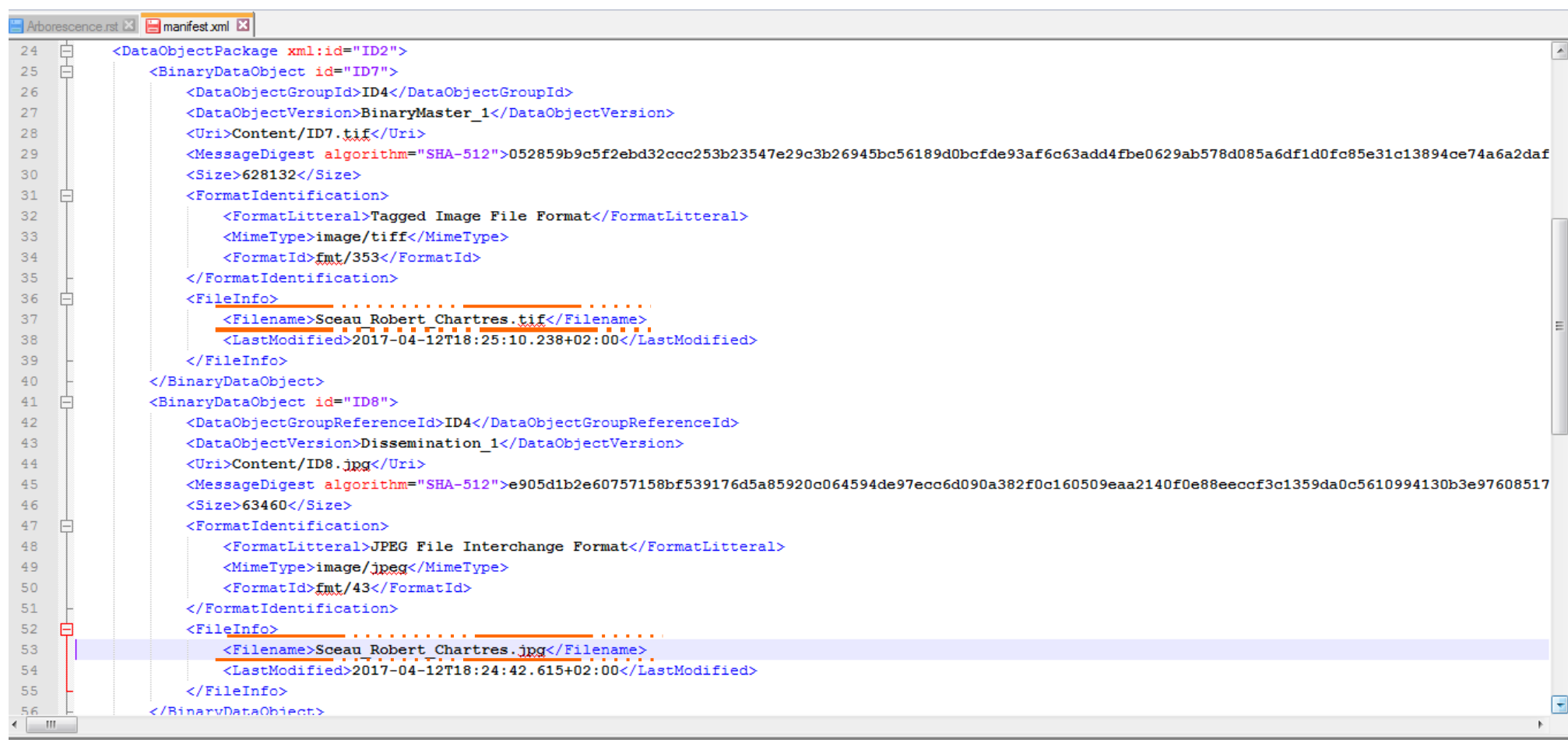
Rassemblement des fichiers constituant des représentations différentes d'une même unité archivistique (2)

- Étape 3 : préfixer le nom de chaque fichier de la manière suivante :
 - 2 Underscores + le type d'usage + 1 Underscore + le numéro de version
 - Exemple :
__BinaryMaster_1_mon nom de fichier.extension
 - Les types d'usages possibles sont :
 - *BinaryMaster* = original numérique
 - *Dissemination* = diffusion
 - *Thumbnail* = vignette
 - *TextContent* = texte brut
- Étape 4 : lancer la création du SIP



Rassemblement des fichiers constituant des représentations différentes d'une même unité archivistique (3)

- Les 2 objets sont regroupés dans un même groupe d'objet dans le bordereau
- Le nom du fichier est le nom d'origine, sans le préfixe rajouté



```
24 <DataObjectPackage xml:id="ID2">
25   <BinaryDataObject id="ID7">
26     <DataObjectGroupId>ID4</DataObjectGroupId>
27     <DataObjectVersion>BinaryMaster_1</DataObjectVersion>
28     <Uri>Content/ID7.tif</Uri>
29     <MessageDigest algorithm="SHA-512">052859b9c5f2ebd32ccc253b23547e29c3b26945bc56189d0bcfde93af6c63add4fbc0629ab578d085a6df1d0fc85e31c13894ce74a6a2daf
30     <Size>628132</Size>
31     <FormatIdentification>
32       <FormatLiteral>Tagged Image File Format</FormatLiteral>
33       <MimeType>image/tiff</MimeType>
34       <FormatId>fmt/353</FormatId>
35     </FormatIdentification>
36     <FileInfo>
37       <Filename>Sceau Robert Chartres.tif</Filename>
38       <LastModified>2017-04-12T18:25:10.238+02:00</LastModified>
39     </FileInfo>
40   </BinaryDataObject>
41   <BinaryDataObject id="ID8">
42     <DataObjectGroupReferenceId>ID4</DataObjectGroupReferenceId>
43     <DataObjectVersion>Dissemination_1</DataObjectVersion>
44     <Uri>Content/ID8.jpg</Uri>
45     <MessageDigest algorithm="SHA-512">e905d1b2e60757158bf539176d5a85920c064594de97ecc6d090a382f0c160509eaa2140f0e88eccc3c1359da0c5610994130b3e97608517
46     <Size>63460</Size>
47     <FormatIdentification>
48       <FormatLiteral>JPEG File Interchange Format</FormatLiteral>
49       <MimeType>image/jpeg</MimeType>
50       <FormatId>fmt/43</FormatId>
51     </FormatIdentification>
52     <FileInfo>
53       <Filename>Sceau Robert Chartres.jpg</Filename>
54       <LastModified>2017-04-12T18:24:42.615+02:00</LastModified>
55     </FileInfo>
56   </BinaryDataObject>
```

Rassemblement des fichiers constituant des représentations différentes d'une même unité archivistique (4)

- L'unité de description est unique
- Elle a pour titre le nom du répertoire, sans les Underscore

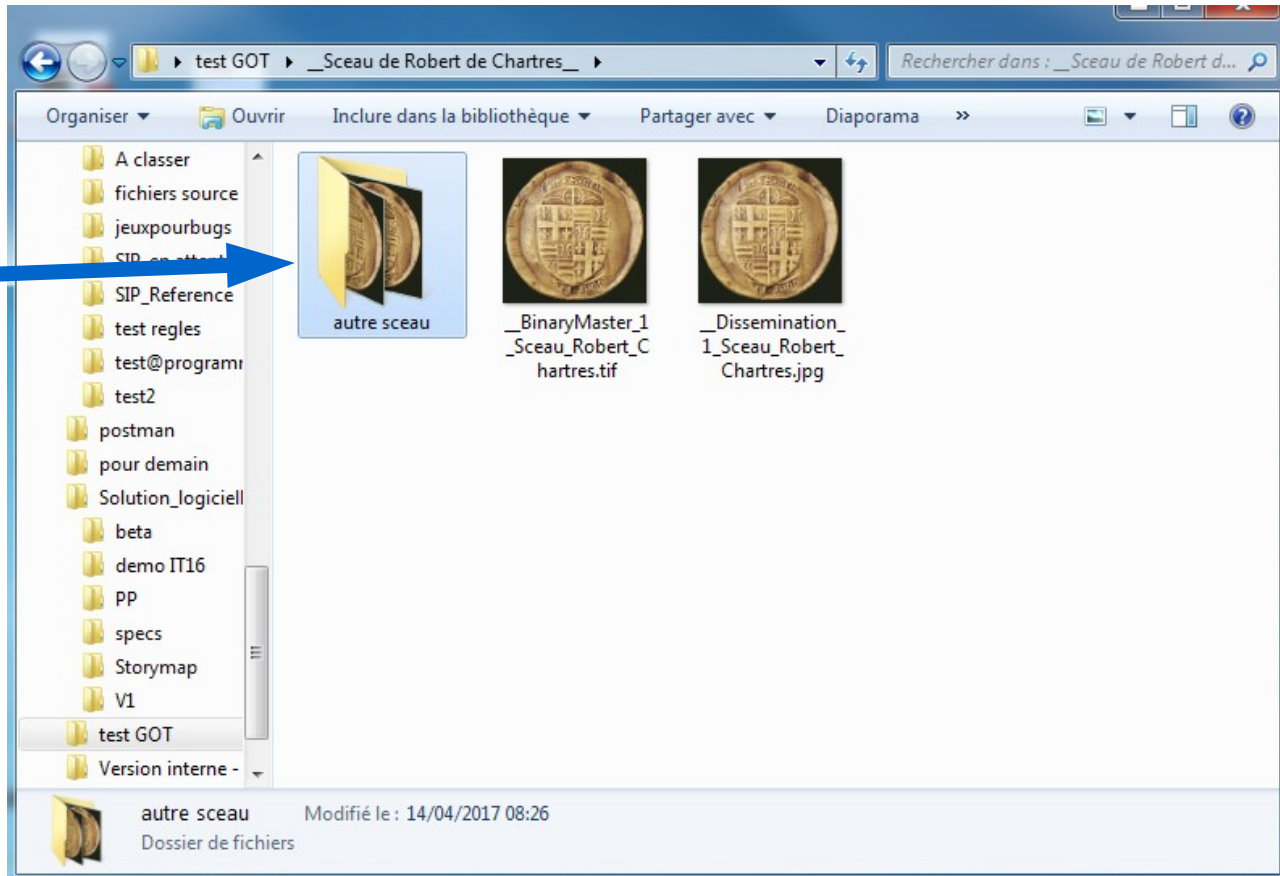
```
</ArchiveUnit>
<ArchiveUnit id="ID5">
  <Content>
    <DescriptionLevel>Item</DescriptionLevel>
    <Title>Sceau de Robert de Chartres</Title>
    <Description>C:\Users\edouard.vasseur\Desktop\test GOT\__Sceau de Robert de Chartres__</Description>
    <TransactedDate>2017-04-12T18:25:10</TransactedDate>
  </Content>
  <DataObjectReference>
    <DataObjectGroupReferenceId>ID4</DataObjectGroupReferenceId>
  </DataObjectReference>
</ArchiveUnit>
</DescriptiveMetadata>
<ManagementMetadata>
  <OriginatingAgencyIdentifier>Service_producteur</OriginatingAgencyIdentifier>
  <SubmissionAgencyIdentifier>Service_versant</SubmissionAgencyIdentifier>
</ManagementMetadata>
</DataObjectPackage>
<ArchivalAgency>
  <Identifier>Identifier4</Identifier>
</ArchivalAgency>
<TransferringAgency>
  <Identifier>Identifier5</Identifier>
</TransferringAgency>
</ArchiveTransfer>
```

Gestion des unités archivistiques complexes, avec fichiers rattachés et également arborescence (1)

- Exemple d'un registre numérisé :
 - À un registre correspond une unité archivistique
 - À cette unité archivistique peut être associé un objet, un fichier multipages
 - Mais cette unité archivistique peut également avoir pour filles d'autres unités archivistiques correspondant à chaque page du registre => elle a donc également une arborescence d'unités archivistiques « filles »
- Comment utiliser le générateur ?
 - Utiliser la méthode décrite pour rassembler des fichiers constituant des représentations différentes d'une même unité archivistique
 - Intégrer dans le répertoire qui a pour préfixe et suffixe les 2 Underscores l'arborescence de répertoires et de fichiers à intégrer

Gestion des unités archivistiques complexes, avec fichiers rattachés et également arborescence (2)


- Étape 1 : rajouter un répertoire dans le répertoire précédemment créé
- Étape 2 : lancer le générateur



Gestion des unités archivistiques complexes, avec fichiers rattachés et également arborescence (3)

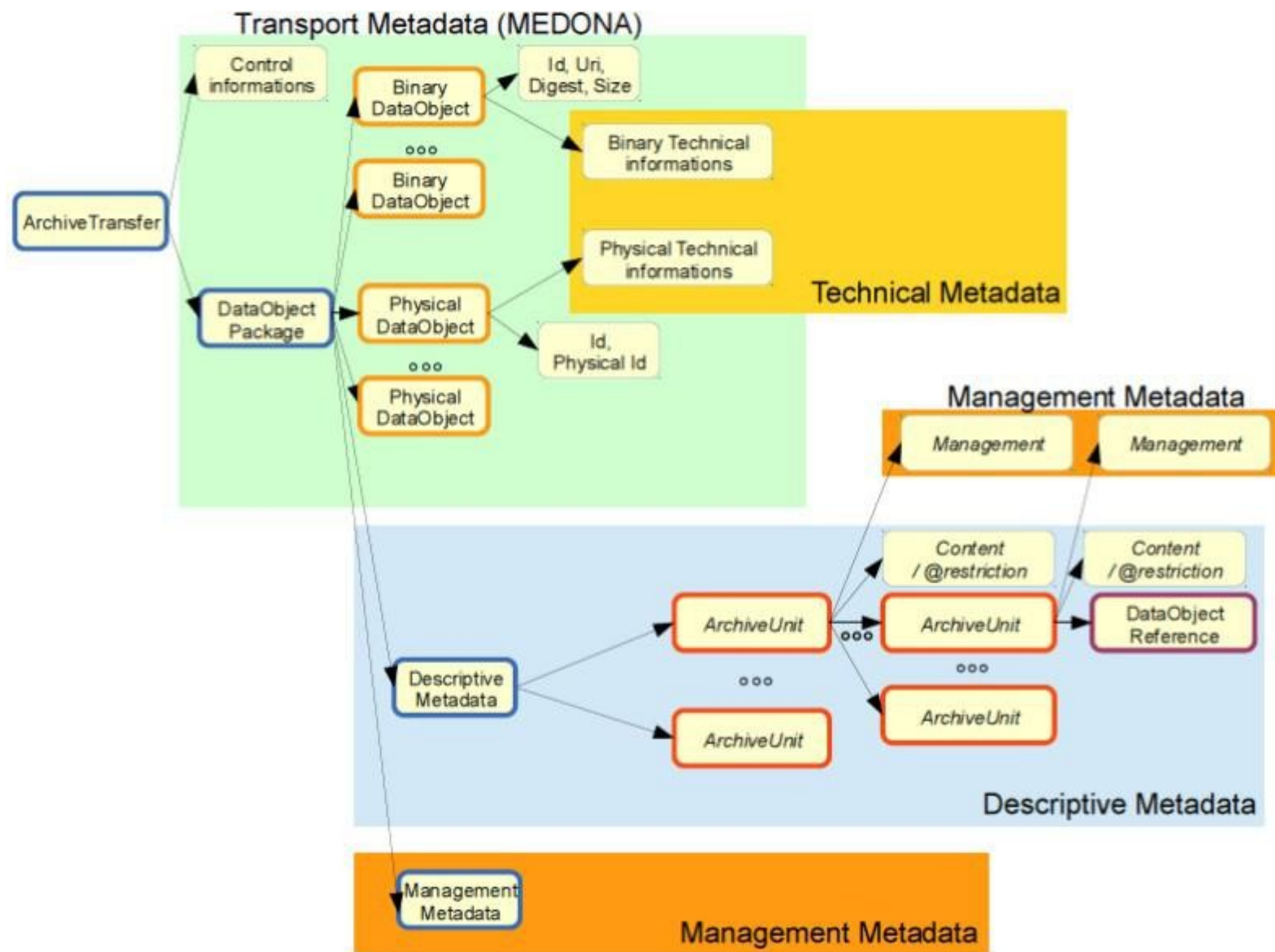
- Dans le manifeste, le répertoire inséré dans le répertoire précédemment créé est devenu une unité archivistique fille de l'unité archivistique correspondant au répertoire

```
</ArchiveUnit>
<ArchiveUnit id="ID5">
  <Content>
    <DescriptionLevel>Item</DescriptionLevel>
    <Title>Sceau de Robert de Chartres</Title>
    <Description>C:\Users\edouard.vasseur\Desktop\test GOT\__Sceau de Robert de Chartres__</Description>
    <TransactedDate>2017-04-12T18:25:10</TransactedDate>
  </Content>
  <DataObjectReference>
    <DataObjectGroupReferenceId>ID4</DataObjectGroupReferenceId>
  </DataObjectReference>
  <ArchiveUnit id="ID8">
    <ArchiveUnitRefId>ID7</ArchiveUnitRefId>
  </ArchiveUnit>
</ArchiveUnit>
<ArchiveUnit id="ID7">
  <Content>
    <DescriptionLevel>RecordGrp</DescriptionLevel>
    <Title>autre sceau</Title>
    <Description>C:\Users\edouard.vasseur\Desktop\test GOT\__Sceau de Robert de Chartres__\autre sceau</Description>
    <StartDate>2017-04-12T18:24:42</StartDate>
    <EndDate>2017-04-12T18:25:10</EndDate>
  </Content>
  <ArchiveUnit id="ID12">
    <ArchiveUnitRefId>ID11</ArchiveUnitRefId>
  </ArchiveUnit>
</ArchiveUnit>
```



Annexe sur le SEDA

Structuration du bordereau (message de type ArchiveTransfer)



Vocabulaire

- <DataObjectPackage>
Englobe tous les objets et leurs MD de description et de gestion
- <BinaryDataObject>
Objet correspondant à des fichiers binaires
- <PhysicalDataObject>
Objet correspondant à quelque chose de physique (un carton, un cd-rom, etc.)
- <DataObjectGroup>
Groupe d'objets données
- <DataObjectGroupId>
Identifiant d'un groupe d'objets données
- <DataObjectVersion>
Usage/version de l'ensemble intellectuel
- <Uri>
Chemin permettant d'accéder au fichier
- <MessageDigest>
Empreinte du fichier
- <Descriptive Metadata>
MD de description
- <Management Metadata>
MD de gestion
- <ArchiveUnit>
Unité d'archives
- <Content>
Contenu de l'unité d'archives
- <DataObjectReference>
Référence interne à un objet-donnée ou à un groupe d'objet-données