Pas à pas : faire un SIP RP PM

Sources qu'il faut préalablement préparer

Préparation des données

Les photos doivent être remises en répertoires (1 répertoire correspondant à un article) en exécutant les commandes Powershell

- Transformation des répertoires présents sur bande LTO (qui regroupent plusieurs reportages) en un répertoire par reportage.
- Voir le tableau de matrice des commandes Powershell pour les exécuter en masse
- Attention, s'il y a des séquences (voir l'IR pour savoir lesquelles), il faut les créer dans l'arborescence cibleavant les analyses DROID

Les CSV de métadonnées DROID

- Importer chaque répertoire année contenant les reportages à traiter dans DROID.
- **Faire un export CSV complet** d'abord pour avoir une vue des formats et repérer les formats problématiques qu'il faudra éliminer (ex : Adobe Photoshop)
- **Faire un export CSV partiel**. Ce deuxième export permettra de préparer un CSV à importer dans ReSIP. Pour celui-ci, choisir quelques champs à exporter : Id, Parent Id, File path, File name, Resource type

L'instrument de recherche

- Convertir l'IR en XML en tableur XLSX grâce à OpenRefine (et le nettoyer : format de dates, positionner la cote des colonnes en première position)

Les exports Exiftool

- Réaliser les exports Exiftool des photos des répertoires formant le SIP à créer, avec les colonnes suivantes : FileName, SourceFile, Creator, By-line, Artist, CreateDate, Keywords, Subject, Location, City, Country, Country-Primary Location Name. L'ouvrir dans Excel pour le nettoyer et l'enregistrer en tableur XLSX
  - Actuellement, le premier export Exiftool est réalisé à partir de la division en dossiers sur bande LTO qui regroupe un ensemble de reportages sans séparation logique/signifiante. Il s'agira ensuite de regrouper le contenu de ces différents csv en fonction des paquets (SIP) que l'on souhaite intégrer dans ReSIP.

Attention : lors de l'export Exiftool, les colonnes de chaque champ ne sont créées que si l'information est présente dans les fichiers analysés. Ex : si aucune des photographies analysées ne contient d'information de localisation, alors les

colonnes « city » / « country » n'apparaitront pas. Cela peut poser problème lorsqu'il s'agira de regrouper les différents CSV Exiftool par année : pour pouvoir Copier/Coller directement l'ensemble d'un CSV et non colonne par colonne, il faut toujours avoir entré la liste des champs dans le même ordre dans les commandes Exiftool, et utiliser la commande **-f** pour forcer Exiftool à créer autant de colonnes qu'il y a de tags demandés : les informations manquantes sont alors signalées par un tiret (-).

- A partir des informations fournies par l'instrument de recherche, rassembler les CSV Exiftool par année. Une année correspondra à 1 à N paquets (SIP).

Nettoyer les exports CSV d'Exiftool

1. Suppression des colonnes inutiles
2. Création de deux csv par année : **un avec les informations conservées au niveau des fichiers** (nom de fichier, date, photographe) et **un autre avec les informations qui seront reportées au niveau de l'article/reportage** (noms de fichiers, localisations, mots-clés)

Attention : lorsque l'on importe un CSV contenant des dates dans Excel, il faut bien s'assurer que le CSV est bien importé au format « Texte » (notamment pour les dates), autrement Excel convertira automatiquement le format des cellules contenant des dates.

Nettoyage du CSV Exiftool niveau « fichiers »

Les dates

- Changer le format de date issu de CreateDate (YYYY:MM:DD HH:MM:SS ===> YYYY-MM-DDTHH:MM:SS)
  - A rechercher : ;([0-9]{4}):([0-9]{2}):([0-9]{2}) ([0-9]{2}):([0-9]{2}):([0-9]{2});
  - Remplacer par : ;$1-$2-$3T$4:$5:$6;
  - Cocher "expression régulière" en bas
- Suppression des dates incohérentes (dates antérieures / postérieures dues à un mauvais réglage des appareils photos)

Les photographes

- Nettoyage manuel :
  - Choisir une colonne dans laquelle on souhaite regrouper les informations des différents champs
  - Trier les colonnes By-line / Creator / Artist sur les vides pour faire ressortir les différences
  - Copier/coller les informations manquantes dans la colonne choisie.
- Utiliser OpenRefine pour uniformiser et corriger les noms des photographes
  - Importer le csv, choisir « ; » comme séparateur

- Sur la colonne « photographe » : « Edit cells » --> « Cluster ans edit… »
- Choisir les formes souhaitées des noms de photographes --> « Merge selected and re-cluster »
- Export « .xlsx »

- Chercher des initiales dans les noms de fichiers pour compléter la métadonnée « photographe » de certains fichiers (si possible)
  - Rechercher : ^[0-9]{6}_[0-9]{8}_[0-9]{4}_[A-Z]+?
  - Commande qui ne fait normalement ressortir que les fichiers dont le nom d'origine commence par une lettre (après nommage Constance). Il s'agira ensuite déterminer s'il s'agit des initiales du photographe ou d'un nommage descriptif.

Nettoyage du CSV Exiftool niveau « article » / « reportage »

Les mots-clés

- En utilisant la même méthode que pour la concaténation des colonnes « photographe », ne conserver qu'une seule colonne « mots-clés » regroupant les informations des colonnes « Subject » et « Keywords ».
- Enregistrer une version du csv avec uniquement les colonnes FileName et Keywords
  - Choisir encodage « Système »
- Ouvrir dans Notepad++

1. Marquer les lignes sans mots-clés puis les supprimer
   - Identification des lignes sans indexation : ^[0-9]{6}_[0-9]{8}_[0-9]{4}_.+?\.(JPG|JPEG);$
   - Identification des lignes des CSV : .csv;
   - Attention : les lignes de fichiers qui ne sont ni des CSV ni des JPG ne seront pas supprimées
   - "Recherche -> Signet -> Supprimer les lignes marquées"
2. Ne conserver que les numéros d'article (sans les extensions/noms de fichiers)
   - RECHERCHER : ^([0-9]{6}_[0-9]{8}_[0-9]{4})(_.+?)(\.JPG|JPEG)
   - REMPLACER : $1
3. Supprimer les lignes doublons
   - Edition --> Ligne --> « Trier les lignes dans l'ordre lexicographique croissant » puis « Supprimer les lignes en double consécutives »
4. Réduire à 1 ligne par mot-clef
   - RECHERCHE : , (virgule espace)
   - REMPLACER : \r\n;
5. Nettoyage supplémentaire
   - RECHERCHE : \r\n (avec espace)
   - REMPLACE : \r\n (sans espace)
6. Copier/Coller dans Excel
   - Données -> Convertir -> Sélectionner le « ; » comme séparateur
   - Si la première ligne est vide, la supprimer
7. Sélectionner les « trous » de la colonne A pour associer le bon numéro d'article à chaque ligne de mots-clés

- Accueil --> Rechercher et sélectionner --> Sélectionner les cellules --> Cellules vides
- Taper « = » (sans rien faire d'autre) ; puis « flèche supérieure » (clavier), puis Ctrl/Entrée

8. Enregistrer le csv
9. Réaliser les traitements OpenRefine de nettoyage et d'alignement

- Importer le csv, choisir « ; » comme séparateur
- Sur la colonne des mots-clés : « Edit cells » --> « Cluster ans edit... »
- Choisir les formes souhaitées des mots clés : « Merge selected and re-cluster »
- Export « .xlsx »

10. Dé-doublonner les lignes dans Excel

- Données -> Supprimer les doublons (sélectionner colonnes A et B)

11. Nettoyage manuel des occurrences uniques ou très mal orthographiées sur Excel
12. Réunir l'ensemble des mots-clés d'un article dans une même cellule

- Ajouter une ligne vide en en-tête
- Saisir "=SI(A2=A1;C1&";"&B2;B2)" dans la cellule C2
- Saisir "=SI(A2<>A3;C2;"")" dans la cellule D2
- Dans les deux cas étendre à toutes les lignes
- Filtrer la colonne D2 sur les non-vides
- Copier/Coller les colonnes A et D dans un nouveau document
  - Si la colonne D est en format Texte, Excel aura automatiquement placé le contenu de chaque cellule entre guillemets droites. Si le CSV est ouvert, même avec le bon séparateur, l'ensemble des mots clé apparaitra au sein d'une même cellule
  - Solution(s) : copier la colonne D en format Standard OU conserver ainsi jusqu'à l'étape finale de réalisation du CSV de métadonnées, puis supprimer les guillemets dans Notepad++

13. Enregistrer le CSV

Les localisations

- Créer un CSV avec les colonnes « FileName », « City », « Country » et « Country-Primary Location Name ».
- Regrouper le contenu des colonnes « Country » et « Country-Primary Location Name ».
- Si toutes les photographies d'un même reportage n'ont pas été prises dans la même ville, ajouter autant de colonnes « City » que nécessaire.
- Nettoyer les noms de lieux mal orthographiés ; traduire les noms de pays qui ont pu être renseignés en anglais.
- Répéter la méthode employée pour les mots-clés afin d'obtenir un CSV avec une ligne par numéro d'article + une colonne par information de localisation.
  - En cas de donnée de pays manquante mais d'indication de ville(s) : a priori, le fait que la première cellule de localisation soit vide ne pose pas de problème à ReSip.

Nota : Il peut être nécessaire de reprendre les mots-clés et localisations manuellement pour compléter le travail réalisé sur OpenRefine, notamment pour

les termes n'ayant qu'une forme incorrecte dans le document, ou les termes (très) mal orthographiés et donc pas repérés par OpenRefine.

Réaliser le CSV de métadonnées à importer de ReSIP

- Dans Excel, ouvrir une nouvelle feuille et importer le CSV de DROID
- Renommer « PARENT_ID » en « ParentID »
- Renommer « FILE_PATH » en « File »
- Le CSV doit contenir que les niveaux reportages. Si besoin, supprimer la ligne qui correspondrait à un niveau parent aux reportages (dossier d'une année par exemple)

Renseigner la balise Content.DescriptionLevel

- Ajouter une colonne « Content.DescriptionLevel »
- Dans la colonne « SIZE », filtrer sur les valeurs non vides pour filtrer sur les lignes correspondant aux fichiers
- Dans la colonne « Content.DescriptionLevel », saisir Item sur la ligne du premier fichier et étirer la valeur sur le reste de la colonne
- Défiltrer « SIZE »
- Filtrer sur « Content.DescriptionLevel » et afficher les lignes vides
- Saisir « RecordGrp » sur la première ligne d'un dossier et étirer la valeur sur le reste de la colonne
- Défiltrer « Content.DescriptionLevel », les fichiers doivent avoir « Item » et les dossiers « RecordGrp »
- Supprimer la colonne « SIZE »

OU

- Identifier le niveau de description
- Créer une nouvelle colonne après "FILE_PATH" et "NAME"
- La nommer "Content.DescriptionLevel"
- Si la cellule de la colonne "SIZE" (ici F) est vide, y écrire RecordGrp, sinon y écrire Item
- Dans la 2ème cellule de la colonne, utiliser la fonction : =SI(ESTVIDE(F2);"RecordGrp";"Item")
- Dupliquer la formule sur toute la colonne.
- Copier/Coller la colonne « Content.DescriptionLevel » en valeur
- Supprimer la colonne « SIZE »

OU

- Utiliser la colonne « Type » de l'export DROID
- Remplacer les « File » par « Item » et les « Folder » par « RecordGrp »

Joindre les informations de l'IR

1. Isoler le numéro d'article (NUMEROVERSEMENT_ARTICLE)

- Insérer une colonne ARTICLE1 à la suite de « File »
- Récupérer le numéro d'article pour les lignes correspondant à RecordGrp en tapant la fonction : =SI(F2="RecordGrp";DROITE(C2;12))
- Etirer la fonction sur le reste de colonne

- Copier/coller la colonne en valeurs
- Attention, le résultat de cette fonction doit être vérifié et corrigé sur les niveaux correspondant à un intervalle de reportages / séquences
- La formule n'extrayant que les 12 derniers caractères, il est nécessaire de corriger manuellement les numéros d'article des répertoires correspondant à un intervalle d'articles (dans la mesure où le numéro à extraire est ici de 16 caractères)

2. Renseigner les intitulés des reportages, leurs dates extrêmes et les anciens numéros d'article (de l'IR)

- Créer une colonne « TITRE REPORTAGE »
- Filtrer « Content.DescriptionLevel » sur « RecordGrp » pour visualiser les niveaux dossiers
- Dans « « TITRE REPORTAGE », sur la première ligne correspondant à un reportage, faire une RECHERCHEV en cherchant ARTICLE1 (qui correspond à la cote) dans l'IR converti en tableur afin de copier le titre du reportage : « =RECHERCHEV(Article1;MatriceIR;PositionColonneUnitTitle;FAUX) »
- Tirer la fonction sur le reste de la colonne « TITRE REPORTAGE »
- Pour les répertoires correspondant à des intervalles d'articles : copier/coller manuellement le titre depuis l'IR
- Créer une colonne « Content.TransferringAgencyArchiveUnitIdentifier »
- Dans « Content.TransferringAgencyArchiveUnitIdentifier », sur la première ligne correspondant à un reportage, reporter la valeur de « ARTICLE1 » par une fonction « =CELLULE » (exemple « =D2 »). Tirer sur le reste de la colonne.
- Créer une colonne « Content.StartDate » et une colonne « Content.EndDate »
- ATTENTION : les cellules doivent être au format TEXTE
- Dans « Content.StartDate », sur la première ligne correspondant à un reportage, faire une RECHERCHEV en cherchant ARTICLE1 (cote) dans l'IR converti en tableur afin de copier les dates extrêmes (ATTENTION : vérifier au préalable que les dates sont au format texte dans le tableur) : « =RECHERCHEV(Article1;MatriceIR;PositionColonneUnitDateNormal;FAUX) . Bien récupérer les dates au format AAAA-MM-JJ.
- Faire la même chose dans « Content.EndDate »
- Défiltrer « Content.DescriptionLevel »
- Copier les colonnes « TITRE REPORTAGE », « Content.TransferringAgencyArchiveUnitIdentifier », « Content.StartDate », « Content.EndDate » et coller les valeurs en texte
- Vérifier les valeurs pour les niveaux correspondant à des intervalles, les corriger si besoin : supprimer la valeur correspondant au TransferringAgencyArchiveUnitIdentifier, renseigner les dates, ajouter le titre à la main.
- Dans les colonnes « Content.StartDate » et « Content.EndDate », nettoyer les valeurs où il y avait des intervalles de dates, afin de les séparer en date de début et date de fin

- Dans « Content.TransferringAgencyArchiveUnitIdentifier », remplacer les « _ » par « / » pour recréer les cotes articles

3. Créer des références aux reportages argentiques

- A la main, si besoin, créer et remplir une colonne « Content.RelatedObjectReference.References.ExternalReference » afin de faire les liens avec les reportages argentiques (cas de JP Raffarin)

Renseigner les métadonnées des photos

- Filtrer la colonne « Content.DescriptionLevel » sur Item afin d'afficher les fichiers

4. Saisir les noms de photographes

- Ouvrir le fichier CSV contenant les dates des photographies et les noms des photographes dans un tableur Excel pour pouvoir utiliser les fonctions RECHERCHEV
- Créer une colonne « Content.AuthorizedAgent.FullName »
- Dans la première ligne de cette colonne, faire une RECHERCHEV à partir de des noms de fichiers se trouvant dans la colonne NAME : « =RECHERCHEV(NAME;ExportExiftoolPhoto;PositionColonnePhotographe;FAUX) »)
- Créer une colonne « Content.AuthorizedAgent.Activity » et une colonne « Content.AuthorizedAgent.Mandate »
- Filtrer sur les valeurs différentes de « 0 » de « Content.AuthorizedAgent.FullName », renseigner « Photographe » dans « Content.AuthorizedAgent.Activity » et quand cela s'applique « Photographe Matignon » dans « Content.AuthorizedAgent.Mandate » (renseigner en tirant les cellules)
- Défiltrer la colonne « Content.AutorizedAgent.FullName » pour réavoir tous les niveaux items (toutes les photos)

5. Saisir les dates des photos

- Filtrer « Content.DescriptionLevel » sur « Item »
- Dans « Content.StartDate », faire un RECHERCHEV sur la première ligne correspondant à une photo : « =RECHERCHEV(NAME;ExportExiftoolPhoto;PositionColonneDate;FAUX) »). Tirer la fonction dans le reste de la colonne
- Faire la même chose pour « Content.EndDate »
- Défiltrer « Content.DescriptionLevel », copier/coller « Content.StartDate », « Content.EndDate » et « Content.AuthorizedAgent.FullName » en valeur.
- Vérifier qu'il n'y ait pas de dates incohérentes (dates antérieures / postérieures dues à un mauvais réglage des appareils photos) et les supprimer le cas échéant.
- Supprimer les valeurs « 0 » dans la colonne « Content.AuthorizedAgent.FullName » et nettoyer si nécessaire.

Renseigner les métadonnées des reportages

1. Isoler le numéro d'article (NUMEROVERSEMENT_0ARTICLE)

- Créer une colonne « ARTICLE2 » en colonne E

- Récupérer le numéro d'article sur 4 chiffres de la colonne « File » quand la ligne décrit un niveau RecordGrp : =SI(G2="RecordGrp";DROITE(GAUCHE(C2;30);13)). Attention la valeur à prendre en compte dans GAUCHE peut changer selon la profondeur de l'arborescence.
- Copier/coller la colonne en valeurs
- Attention, le résultat de cette fonction doit être vérifié et corrigé sur les niveaux correspondant à un intervalle de reportages / séquences
- S'assurer que la forme de la cote article dans le(s) CSV avec les mots-clés et les localisations est similaire à celui de la colonne ARTICLE2 (il peut être nécessaire de supprimer le numéro du NP en préfixe si cela n'a pas été fait lors de la création du CSV)

2. Importer les métadonnées de localisation

- Dans Excel, ouvrir les données du fichier de localisation (supprimer le préfixe de NP si besoin dans la première colonne)
- Créer une colonne « Content.Coverage.Spatial.1 » (pour le pays) et « Content.Coverage.Spatial.2 » (pour la ville). S'il y a plus d'une ville associée à un même article, ajouter d'autres colonnes à la suite (Coverage.Spatial.3, 4, 5…).
- Filtrer sur « Content.DescriptionLevel » = RecordGrp
- Dans les colonnes Coverage.Spatial, faire une RECHERCHEV en cherchant ARTICLE2 (qui correspond à la cote) dans le CSV contenant les métadonnées de localisation : « =RECHERCHEV(Article2;MatriceLocalisation;PositionColonnePAYS;FAUX) » et « =RECHERCHEV(Article2;MatriceLocalisation;PositionColonneVILLE;FAUX) »

3. Importer les mots-clés

- Dans Excel, ouvrir les données du fichier de mots-clés (supprimer le préfixe de NP si besoin dans la première colonne)
- Créer une colonne « Content.Tag.1 » en dernière position du CSV de métadonnées
- Faire une RECHERCHEV en cherchant ARTICLE2 (qui correspond à la cote) dans le CSV contenant les mots-clés : « =RECHERCHEV(Article2;MatriceMotsClés;PositionColonneUnitTitle;FAUX) »
- L'ensemble des mots-clés sera importé dans une même colonne
- Copier/Coller les colonnes Coverage.Spatial et Tag en valeurs (nécessite de défiltrer la colonne Content.DescriptionLevel) et supprimer les valeurs #N/A et 0
- Si les mots-clés sont importés avec des guillemets encadrants, veillez à ce qu'il n'y ait plus de valeurs sous forme de fonction (sinon copier/coller en valeur) et enregistrer au format CSV avec séparateur point-virgule et ouvrir le CSV de métadonnées dans Notepad++ pour supprimer les guillemets droits empêchant la séparation des mots-clés en plusieurs colonnes

- Si nécessaire, changer préalable le séparateur de la colonne « Content.Tag.1 pour qu'il corresponde au signe de ponctuation utilisé entre les mots-clés (sinon virgule)
- Ajouter un titre à chaque colonne contenant des mots-clés (Content.Tag.2, 3, 4…) en tirant vers la colonne la plus à droite contenant une valeur.

Anticiper le tri alphabétique de Resip : ajouter le numéro de reportage aux titres des reportages

- Créer une colonne « Content.Title » en colonne F
- Si Description.Level = « RecordGrp », concaténer ARTICLE2 + "$$$" + TITRE REPORTAGE, sinon copier NAME : =SI(Content.DescriptionLevel="RecordGrp";CONCATENER(ARTICLE2; "$$$";TITRE REPORTAGE);NAME). Exemple : =SI(H2="RecordGrp";CONCATENER(E2;"$$$";I2);G2)
- Tirer la fonction sur la colonne « Content.Title »
- Copier/Coller en valeur la colonne « Content.Title »
- Pour les répertoires correspondant à un intervalle d'articles, reprendre le numéro de versement et le numéro d'article de l'intervalle.

Finaliser le CSV de métadonnées avant import dans ReSIP

- Vérifier qu'il n'y a plus de colonne contenant des fonctions, mais bien les résultats en valeurs
- Supprimer les colonnes en trop : "ARTICLE1", "ARTICLE2", "NAME", "SIZE", "TITRE REPORTAGE". Ne garder que : ID, ParentID, File, Content.DescriptionLevel, Content.Title, Content.TransferringAgencyArchiveUnitIdentifier, Content.StartDate, Content.EndDate, Content.AuthorizedAgent.FullName, Content.AuthorizedAgent.Activity, Content.AuthorizedAgent.Mandate, les Content.Coverage.Spatial.1 et suivants, les Content.Tag.1 et suivants.
- Vérifier dans les intitulés de colonnes qu'il n'y a pas d'espace à la fin
- Exporter en CSV (séparateur « ; »)

Importer le CSV de métadonnées dans Resip

- Importer le CSV dans ReSip (faire attention au séparateur et à l'encodage dans les Préférences d'import (« ; » et Windows-1252)
- Vérifier qu'il n'y a pas eu de problème d'import de fichiers (le contrôler par la fonctionnalité « Nettoyer les inutiles »)
- Trier l'arbre de visualisation et régénérer les ID continus
- Exporter le SIP

Nettoyer les intitulés de reportages dans le manifest

- Ouvrir le manifest dans Notepad++
- Supprimer l'ajout du préfixe utilisé que pour le classement

- Rechercher : <Title>[0-9]{8}_[0-9]{4}\$\$\$
- Remplacer : <Title>
- Enregistrer sous le nouveau manifest