

# **Outils et méthodologie pour le tri archivistique de supports de données complexes**

**Travail de master réalisé par :  
Denis BUSSARD**

Sous la direction de :  
**Frédéric SARDET, directeur de la Bibliothèque de Genève**

**Lausanne, 15 août 2022**

**Information documentaire  
Haute École de Gestion de Genève (HEG-GE)**

## Résumé

Ce Mémoire de Master est le fruit d'un mandat réalisé pour le compte de la Cinémathèque suisse portant sur l'évaluation et le tri archivistiques de supports de données complexes. Ce que nous appelons fréquemment « vrac numérique » ou « données non-structurées » pose effectivement des problèmes spécifiques pour les institutions patrimoniales qui les retrouvent en grand nombre dans les fonds d'archives privées qu'elles conservent. Une fois ces données numériques extraites de leurs supports physiques (clés USB, disques durs externes, etc.), et avant leur archivage à long terme sur des bandes magnétiques hautement sécurisées, il convient de *traiter* ces masses gigantesques de données. La première tâche consiste à évaluer les contenus pour décider de leur sort final : quels sont les documents dont la valeur patrimoniale justifie une conservation à long terme et quels sont les dossiers et fichiers qui pourront être éliminés.

La masse très conséquente de données et la difficulté à appréhender les arborescences de fichiers créées par des personnes privées rendent nécessaire l'utilisation d'outils informatiques permettant de prendre connaissance des contenus des supports de données, mais aussi d'automatiser certaines tâches (comme la recherche de redondances strictes, qui ne peut être réalisée manuellement). Cette recherche vise donc non seulement à identifier les logiciels actuellement disponibles et utilisables par des archivistes sans formation informatique avancée, mais propose aussi de définir une liste d'étapes, de tâches à réaliser pour mener à bien l'évaluation et le tri archivistiques des supports de données complexes au niveau macroscopique (l'évaluation des contenus au niveau des fichiers étant manuellement irréaliste).

Onze logiciels ont été testés via une grille d'analyse standardisée et une recommandation d'acquisition est formulée. Les outils les plus utiles sont les suivants : *Archifiltre*, *AllDup*, *AntiDupl*, *Beyond Compare*, *Droid*, *Karen's Directory Printer* et *TreeSize Professional*. Aucun outil ne remplissant à lui seul toutes les attentes et tous les besoins de la Cinémathèque en termes de traitement, une utilisation combinée et parallèle de ces outils sera nécessaire.

Cinq tâches ont en outre pu être identifiées : l'extraction de métadonnées et l'établissement d'un récolement ; l'analyse de l'arborescence ; le traitement des dossiers vides ; le traitement des redondances strictes ; la comparaison de données. Chacune de ces tâches est ensuite décomposée en sous-tâches, en points d'attention (soit des analyses permettant la prise de décision), en traitements possibles et en une liste d'outils permettant de les réaliser. Les fonds d'archives de Francis Reusser et d'Ana Simon, mis à disposition par la Cinémathèque suisse, ont servi à la réalisation d'études de cas menées grâce aux outils sélectionnés et sur la base de propositions d'analyse et d'instruments méthodologiques supplémentaires.

Mots-clés : archivage numérique ; évaluation ; méthodologie de tri ; supports de données complexes ; logiciels d'archivage.

# Table des matières

<b>Résumé .....</b>	<b>i</b>
<b>Liste des tableaux .....</b>	<b>iv</b>
<b>Liste des figures.....</b>	<b>v</b>
<b>1. Introduction.....</b>	<b>1</b>
<b>1.1 Problématique .....</b>	<b>1</b>
<b>1.2 Contexte institutionnel et archivistique : la Cinémathèque suisse .....</b>	<b>3</b>
1.2.1 Histoire et mandat de collection .....	3
1.2.2 Organisation et projets numériques.....	4
1.2.3 Archivage des données numériques à la Cinémathèque suisse : le projet <i>Ingest Manager</i> .....	6
<b>1.3 Revue de littérature.....</b>	<b>9</b>
1.3.1 Projets précédents et méthodologies existantes .....	9
1.3.2 Listes d'outils .....	12
<b>1.4 Objectifs et questions de recherche.....</b>	<b>13</b>
<b>2. Méthodologie générale.....</b>	<b>15</b>
<b>2.1 Typologie de la recherche .....</b>	<b>15</b>
<b>2.2 Échantillon.....</b>	<b>15</b>
<b>2.3 Périmètre de la recherche .....</b>	<b>17</b>
<b>3. Analyse des fonctionnalités logicielles .....</b>	<b>19</b>
<b>3.1 Méthodologie.....</b>	<b>19</b>
3.1.1 Critères de sélection des fonctionnalités .....	19
3.1.2 Critères de sélection des outils testés .....	20
3.1.3 Liste des outils testés.....	20
3.1.4 Grille d'analyse des fonctionnalités et des logiciels .....	34
<b>3.2 Objectifs : une grille d'analyse pérenne et une base d'évaluation sûre..</b>	<b>36</b>
<b>3.3 Résultats.....</b>	<b>36</b>
3.3.1 Extraction de métadonnées / Récolement.....	36
3.3.2 Arborescence et volumétrie.....	48
3.3.3 Recherche de redondances strictes .....	61
3.3.4 Comparaison de données .....	74
<b>3.4 Discussion méthodologique .....</b>	<b>84</b>
3.4.1 Formulation et abstraction.....	84
3.4.2 Autonomie et dépendance des micro-fonctionnalités et informations ....	84
3.4.3 Accomplissement intégral ou partiel d'une fonctionnalité.....	84
<b>4. Méthodologie de tri archivistique .....</b>	<b>86</b>
<b>4.1 Extraction de métadonnées / Récolement.....</b>	<b>87</b>
<b>4.2 Analyse de l'arborescence .....</b>	<b>88</b>

4.2.1	Analyse de la profondeur de l'arborescence.....	88
4.2.2	Analyse de la structure de l'arborescence .....	92
4.2.3	Analyse du « plan de classification » initial.....	101
<b>4.3</b>	<b>Traitement des dossiers vides .....</b>	<b>105</b>
<b>4.4</b>	<b>Traitement des redondances strictes .....</b>	<b>107</b>
4.4.1	Traitement des redondances : aperçu général .....	107
4.4.2	Traitement des redondances : analyse des répertoires parents .....	110
4.4.3	Comprendre les relations inter-redondances : essais méthodologiques 112	
4.4.4	Quels modèles ( <i>pattern</i> ) pour les redondances ? .....	119
4.4.5	Le travail de sélection des redondances commence .....	122
<b>4.5</b>	<b>Comparaison de données .....</b>	<b>124</b>
4.5.1	Comparer des répertoires .....	124
4.5.2	Comparer les métadonnées des fichiers .....	130
4.5.3	Comparaison d'images via des algorithmes .....	134
<b>5.</b>	<b>Conclusion .....</b>	<b>138</b>
	<b>Bibliographie .....</b>	<b>140</b>



## Liste des tableaux

Tableau 1 : « Extraction de métadonnées / Récolement », par métadonnées / informations, résumé.....	38
Tableau 2 : « Extraction de métadonnées / Récolement », par outils, résumé.....	41
Tableau 3 : Cartographie des métadonnées extraites .....	45
Tableau 4 : Tableau analytique des métadonnées extraites .....	46
Tableau 5 : Tableau analytique « Extraction de métadonnées / Récolement », qualité générale .....	47
Tableau 6 : « Arborescence et volumétrie », par informations et micro-fonctionnalités, résumé .....	49
Tableau 7 : « Arborescence volumétrie », par outils, résumé .....	50
Tableau 8 : Tableau analytique « Arborescence et volumétrie » .....	60
Tableau 9 : « Recherche de redondances strictes », par informations et micro-fonctionnalités, résumé .....	63
Tableau 10 : « Recherche de redondances strictes », par outils, résumé .....	64
Tableau 11 : Tableau analytique « Recherche de redondances strictes » .....	73
Tableau 12 : Extraction de métadonnées / Récolement, tâches .....	87
Tableau 13 : Arborescence et volumétrie : cas pratiques, résumé .....	88
Tableau 14 : Analyse de la structure de l'arborescence, Fonds Simon .....	95
Tableau 15 : Analyse de la structure de l'arborescence, Fonds Reusser .....	96
Tableau 16 : Statistiques relatives au nombre de fichiers à la racine, Fonds Reusser .....	98
Tableau 17 : Distribution des fichiers à la racine par dossier, Fonds Reusser .....	99
Tableau 18 : Nomenclature des dossiers, niveau 4, Fonds Simon .....	103
Tableau 19 : Arborescence et volumétrie, tâches .....	104
Tableau 20 : Traitement des dossiers vides, tâches .....	105
Tableau 21 : Recherche de redondances strictes : cas pratiques, résumé .....	107
Tableau 22 : Éléments non-uniqes au sein d'un répertoire, Fonds Reusser .....	112
Tableau 23 : Fichier source pour la base de données relationnelle, dossiers redondants, Fonds Reusser .....	113
Tableau 24 : Nombre de dossiers redondants par répertoire de niveau 2 et 3, et mesure de dispersion, Fonds Reusser .....	115
Tableau 25 : « Base de données relationnelle » pour identifier les liens entre répertoires parents, dossiers redondants, Fonds Reusser .....	116
Tableau 26 : Analyse des embranchements divergents, fichiers redondants, Fonds Reusser .....	118
Tableau 27 : Traitement des redondances strictes, tâches .....	123
Tableau 28 : Exemples de fichiers redondants de différentes extensions, Fonds Reusser .....	130
Tableau 29 : Comparaison de données, tâches .....	137

## Liste des figures

Figure 1 : Organigramme de la Cinémathèque suisse.....	5
Figure 2 : Entités fonctionnelles OAIS.....	6
Figure 3 : « Flux numérique, Archivage / INGEST, Non-Film ».....	8
Figure 4 : « Flux de traitements des lots de données au Département Non-Film ».....	9
Figure 5 : Quelques-unes des clés USB d'Ana Simon, Fonds Simon.....	16
Figure 6 : Familles de logiciels par fonctionnalité.....	21
Figure 7 : Types de logiciels (caractéristiques et développements).....	22
Figure 8 : Interface graphique principale de <i>Karen's Directory Printer</i> , Fonds Simon ..	42
Figure 9 : Récolement proposé par <i>Karen's Directory Printer</i> , Fonds Reusser.....	43
Figure 10 : Vue « Extensions », avec les fichiers « .doc » surlignés dans l'arborescence, <i>TreeSize</i> , Fonds Reusser.....	51
Figure 11 : Vue « Détails » avec les éléments classés en fonction du nombre de dossiers, <i>TreeSize</i> , Fonds Reusser.....	52
Figure 12 : Interface unique de <i>WinDirStat</i> , Fonds Reusser.....	53
Figure 13 : Onglet « Général » d' <i>Archifiltre</i> , classement et pondération par volume, Fonds Reusser.....	55
Figure 14 : Onglet « Audit » d' <i>Archifiltre</i> sur les types de fichiers, Fonds Reusser.....	56
Figure 15 : Interface principale de <i>Droid</i> , Fonds Reusser.....	57
Figure 16 : Rapports proposés par <i>Droid</i> .....	58
Figure 17 : Affichage des résultats filtrés, extension « .doc », <i>Droid</i> , Fonds Reusser ..	59
Figure 18 : Recherche de redondances strictes, et filtres de sélection prédéfinis, <i>AllDup</i> , Fonds Simon.....	65
Figure 19 : Recherche de redondances strictes, <i>TreeSize</i> , Fonds Reusser.....	67
Figure 20 : Fonctionnement de la « déduplication », par la création de liens durs ( <i>hardlinks</i> ).....	67
Figure 21 : Onglet « Redondances », <i>Archifiltre</i> , Fonds Reusser.....	69
Figure 22 : « Paramètres de session » pour la comparaison de dossiers, <i>Beyond Compare</i> .....	76
Figure 23 : Affichage des résultats de la comparaison entre deux répertoires (« ROUSSEAU_2011 » et « ROUSSEAU_2012 »), Fonds Reusser, <i>WinMerge</i> .....	77
Figure 24 : Légendes des couleurs pour la comparaison de dossiers et affichages sélectifs des résultats, <i>Beyond Compare</i> .....	78
Figure 25 : Affichage des résultats de la comparaison entre deux répertoires, <i>Beyond Compare</i> , Fonds Reusser.....	78
Figure 26 : Menu « Options » (avec les valeurs par défaut), <i>AntiDupl</i> .....	80
Figure 27 : Résultats de recherche d'images similaires, <i>AntiDupl</i> , Fonds Reusser.....	81
Figure 28 : Nombre de dossiers par niveau de répertoire, Fonds Simon.....	89
Figure 29 : Nombre de dossiers par niveau de répertoire, Fonds Reusser.....	90
Figure 30 : Nombre de dossiers finaux non vides par niveau, Fonds Reusser.....	90
Figure 31 : Nombre total de fichiers par niveau, Fonds Reusser.....	91
Figure 32 : Onglet « Général », <i>Archifiltre</i> , classement par volume et pondération par nombre, Fonds Reusser.....	92
Figure 33 : Onglet « Général », <i>Archifiltre</i> , classement et pondération par volume, Fonds Simon.....	93
Figure 34 : Onglet « Audit », <i>Archifiltre</i> , classement par volume et pondération par nombre, Fonds Simon.....	94
Figure 35 : Nombre de dossiers finaux non vides par niveau, Fonds Simon.....	95
Figure 36 : Dossiers cylindriques, Fonds Reusser, hfsplus\aaCINEATELIER\aaROUSSEAU\ROUSSEAU_2012\IMAGES\PHOTOS_NH97	
Figure 37 : Fichiers à la racine et regroupements thématiques, <i>WinCatalog</i> , Fonds Reusser.....	100

Figure 38 : Fichiers à racine et dossiers, exemple de co-présence, <i>WinCatalog</i> , Fonds Reusser.....	101
Figure 39 : Développement sélectif de l'arborescence, niveau 4, <i>TreeSize</i> , Fonds Simon.....	102
Figure 40 : Dossiers vide et non-vide, « LE TAILLANDIER », <i>WinCatalog</i> , Fonds Reusser.....	106
Figure 41 : Nombre d'occurrences de chaque dossier redondant, par groupe, Fonds Reusser.....	109
Figure 42 : Nombre de fichiers contenus dans un dossier redondant, Fonds Reusser.....	110
Figure 43 : Dossiers redondants possédant le même répertoire parent, <i>TreeSize</i> , Fonds Reusser.....	110
Figure 44 : Nombre de dossiers redondants par répertoire parent, Fonds Reusser.....	111
Figure 45 : Modèles de redondances, emboîtement, Fonds Reusser.....	119
Figure 46 : Modèles de redondances, reports chronologiques, Fonds Reusser.....	120
Figure 47 : Indices de l'utilisation du support : lettres « aa » dans l'intitulé des répertoires, Fonds Reusser.....	121
Figure 48 : Structure du répertoire « ROUSSOPOULOS », pour comparaison de dossiers, Fonds Reusser.....	125
Figure 49 : Comparaison de répertoires, contrôle d'emboîtement exact, <i>Beyond Compare</i> , Fonds Reusser.....	126
Figure 50 : Comparaison de répertoires, reports chronologiques, <i>Beyond Compare</i> , Fonds Reusser.....	127
Figure 51 : Comparaison de répertoires, « méta-classement », éléments identiques (donc reportés), <i>Beyond Compare</i> , Fonds Reusser.....	128
Figure 52 : Comparaison de répertoires, « méta-classement », éléments orphelins à gauche (non reportés), Fonds Reusser, <i>Beyond Compare</i> .....	128
Figure 53 : Statistiques de comparaison de dossiers « A CLASSER\ROUSSEAU » et « aaCINEATELIER\ROUSSEAU », <i>Beyond Compare</i> , Fonds Reusser.....	129
Figure 54 : Comparaison de répertoires généraux / sélectifs (« MARASCO COMPLET » versus « choix pour francis »), <i>Beyond Compare</i> , Fonds Reusser.....	129
Figure 55 : Proportion des intitulés identiques de fichiers, par extension, Fonds Reusser.....	131
Figure 56 : Comparaison de métadonnées, « Nom sans extension », <i>TreeSize</i> , Fonds Reusser.....	132
Figure 57 : Comparaison de fichiers texte, avec toutes les différences, <i>Beyond Compare</i> , Fonds Reusser.....	133
Figure 58 : Comparaison de fichiers texte, différences mineurs ignorées, <i>Beyond Compare</i> , Fonds Reusser.....	134
Figure 59 : Comparaison d'images similaires, exemple d'échec, <i>AllDup</i> , Fonds Simon.....	136
Figure 60 : Comparaison d'images similaires, exemple de réussite, <i>AllDup</i> , Fonds Simon.....	136

# 1. Introduction

## 1.1 Problématique

Cinq mille. C'est le nombre actuel de supports de données numériques qui ont été identifiés dans les collections de la Cinémathèque suisse lors d'une première enquête (François, Rochat 2022). Avec une cinquantaine de fonds d'archives reçus par le Département Non-Film de la Cinémathèque pour les seules années 2020 et 2021 (Cinémathèque suisse 2020a; 2021), et l'utilisation généralisée des ordinateurs depuis l'avènement de l'informatique conviviale à la fin du siècle dernier, il ne fait aucun doute que les supports numériques vont continuer d'affluer en masse au Centre de recherche et d'archivage de Penthaz. Et encore : le nombre total de supports physiques de données (CD-DVD, disquettes, clés USB ou disques durs externes) ne dit rien de la masse documentaire « dématérialisée » que cela représente : petits et opaques, ces supports contiennent en réalité des milliers de dossiers et de fichiers numériques. Si l'extraction de ces données s'impose très rapidement, entre autres pour des questions de préservation relatives à l'obsolescence des supports et de leurs lecteurs, la question de leur traitement *post-extraction* reste entière : que faire et comment traiter une telle masse de données ?

Dans le domaine archivistique, la croissance exponentielle de la masse documentaire induite notamment par le numérique est depuis longtemps un sujet de préoccupation et de réflexion : alors que certains s'appuient sur l'augmentation des capacités de stockage (et la baisse des coûts), mais aussi sur les algorithmes de recherche qui favorisent la découverte dans des océans de données structurées et non-structurées pour défendre une conservation intégrale (Gaudinat 2016), d'autres en revanche s'inquiètent de cette « surabondance informationnelle » (Coutaz 2016) et placent au centre de leur pratique et de la déontologie professionnelle la tâche d'évaluer les archives, tout en insistant sur la pertinence encore actuelle du catalogage et sur la nécessité de continuer à proposer des « instruments de recherche » (que ne remplaceront donc pas les métadonnées) (Langdon 2016). Ce Mémoire de Master, sous la forme d'un mandat proposé par la Cinémathèque suisse à la Haute École de Gestion de Genève, dans le but de proposer une « méthodologie pour le tri archivistique des supports de données complexes », se place donc résolument du côté des tenants de l'évaluation archivistique et de la sélection des documents numériques à conserver. Pour des raisons historiques et archivistiques tout d'abord comme en témoigne la *Politique de Collection* de l'institution :

*« La notion d'"élimination" a longtemps été taboue et même combattue à la Cinémathèque suisse. Mais depuis quelques années, la conscience archivistique s'est développée et l'importance d'éliminer pour offrir une collection (film et non-film) contrôlée, accessible et valorisée, est maintenant considérée comme primordiale. »*  
(Cinémathèque suisse 2015a, p. 29)

Pour des raisons économiques également puisque « toute acquisition, même si elle est réalisée à titre gratuit (don, donation), implique pour l'institution un coût : en effet, il faut traiter le fonds, l'enregistrer, le classer, l'inventorier, le conditionner, le communiquer, le mettre en valeur, le conserver à long terme... » (Roth-Lochner, Gisler 2007, p. 308).

Pour des raisons écologiques enfin : la salle réfrigérée, au sous-sol de la Cinémathèque, qui héberge les bandes *LTO* (*Linear Tape-Open*, sur lesquelles sont stockées les données) dans

d'énormes serveurs clignotants, rappelle que la dématérialisation est un concept discutable et que la conservation à long terme ne va pas sans coûts énergétiques.

Fort de ces prérequis, la présente étude cherchera des moyens d'évaluer et de trier les documents numériques reçus par la Cinémathèque sur des supports de données. Trier plus qu'évaluer d'ailleurs si l'on se fie à la distinction qui existe entre ces deux termes dans l'archivistique contemporaine : si l'évaluation est souvent définie comme « l'acte de juger des valeurs que présentent les documents d'archives (valeur primaire et valeur secondaire) » (Couture 1999, p. 104), acte qui peut avoir lieu à différentes étapes du cycle de vie des documents, le tri « s'apparente surtout à une opération qui intervient en cours de traitement des archives définitives par l'archiviste » (Ducharme 2001, p. 20), une « opération consistant à séparer, dans un ensemble de documents, ceux qui doivent être conservés de ceux qui sont destinés à être détruits » (Portail International Archivistique Francophone (PIAF) 2015a). Dans le cadre de cette étude, il s'agira bel et bien de traiter des archives qui ont *déjà* été acquises par la Cinémathèque parce qu'elles répondaient aux critères énoncés dans la *Politique de Collection* ; et, surtout, il ne s'agira pas d'*évaluer* les archives selon des critères prédéfinis, relatifs à leur valeur informationnelle ou patrimoniale, mais bien d'étudier quelles sont les « opérations » et les tâches qu'il convient de réaliser pour *pouvoir décider* du sort final des documents.

Si le traitement des documents numériques s'appuie sur les mêmes concepts archivistiques fondamentaux (l'acquisition, l'évaluation, le respect du principe de provenance, l'importance accordée au contexte, etc.), leur nature particulière nécessite des moyens et des procédures différents de ceux dont on avait l'habitude avec les collections analogiques. Leur masse rend effectivement impossible et irréaliste une évaluation au niveau du document individuel ; le traitement devra donc adopter une approche *macroscopique* et *descendante* (soit au niveau des répertoires, des séries et des dossiers et non au niveau des seuls fichiers). Le concept « *More Product, Less Process* » (Greene, Meissner 2005), qui a essaimé dans le domaine archivistique depuis une vingtaine d'années, peut aussi s'appliquer à l'évaluation des documents numériques (bien qu'il ait été formulé avant tout pour la description et le catalogage des collections analogiques) : il s'agit de réduire le nombre de manipulations et d'actions ainsi que le niveau de détail du traitement de sorte à pouvoir mettre les collections à la disposition des chercheurs et chercheuses le plus rapidement possible. Pour cela, et grâce à l'informatique, il existe des outils pour faciliter et accélérer l'évaluation de masses gigantesques de données : l'automatisation des tâches archivistiques est en bonne voie et devrait permettre non seulement de diminuer le passif non-traité qui s'accumule dans les institutions (Trace 2021), mais aussi de réduire le nombre de tâches répétitives et d'éliminer des sources d'erreurs potentielles (Morisod 2018). Il s'agira donc dans ce Mémoire de Master d'étudier quelques-unes de ces « possibilités » informatiques via l'analyse d'une dizaine de logiciels permettant de faciliter et d'automatiser certaines tâches (on pense par exemple au traitement des éléments redondants à l'intérieur d'un fonds d'archives).

Outre la volumétrie, l'une des principales difficultés auxquelles sont confrontées les institutions en charge du patrimoine culturel – qui conservent entre autres des archives privées, non soumises à une obligation légale de dépôt –, c'est la complexité des plans de classement des documents : non intégrés dans un système de *Gestion Électronique des Documents* (GED) ou dans une base de données, et n'ayant pas fait l'objet de procédures et de suivi dès leur création (par l'ajout de métadonnées normalisées), ces documents sont effectivement organisés et structurés selon des logiques propres aux producteurs des documents. Ce que

l'on appelle en français « vrac numérique » ou « données non-structurées », s'intitule dans le monde germanophone *Dateiablage* ou *Dateisammlungen* que l'on pourrait définir ainsi en suivant la proposition formulée lors d'un colloque ayant eu lieu en 2016 à Munich et intitulé *Kreative digitale Ablagen und die Archive* :

« *Daten in Dateisammlungen, mithin: Einzeldateien in Filesystemen oder E-Mail-Postfächer, die ohne maschinell nachvollziehbare Struktur abgelegt werden und ohne das Wissen des Anlegenden oft kryptisch bleiben* » (Miegel, Schieber, Schmidt 2017, p. 7)

Ces « supports de données complexes » (ou seulement « supports de données » dans la suite de cette étude) présentent enfin la particularité de n'avoir généralement pas fait l'objet de tri préalable de la part des donateurs : l'acquisition se faisant principalement par des dons, il est effectivement difficile d'exiger des créateurs que leurs archives soient expurgées de tous les documents sans valeur patrimoniale. Cette tâche incombe donc à l'institution et aux archivistes en charge des collections. Nous adopterons donc dans ce travail leur point de vue, en choisissant une approche très « pratique » et « concrète » : que doit-on faire et que peut-on faire avec ces supports de données complexes en termes de tri ? Et avec quels outils informatiques les aborde-t-on ?

## **1.2 Contexte institutionnel et archivistique : la Cinémathèque suisse**

### **1.2.1 Histoire et mandat de collection**

Fondée en 1948 par les membres du Ciné-club de Lausanne, la Cinémathèque suisse est issue des Archives cinématographiques suisses (*Schweizer Filmarchiv*), qui durent fermer leurs portes faute de moyens cinq ans seulement après leur création en 1943 à Bâle. Suite au transfert des collections à Lausanne (1949), la Cinémathèque suisse connaît une augmentation massive de ses activités et une professionnalisation croissante, sous l'impulsion notamment de Freddy Buache (1924-2019), figure désormais associée étroitement à l'institution qu'il dirigea de 1951 à 1996. En 1981, la Cinémathèque, soutenue par le Canton de Vaud (dès 1955) et par la Confédération (à partir de 1963), devient une fondation privée d'utilité publique, avec à sa tête un conseil de fondation composé de neuf membres (présidé actuellement par Jean Studer) et un directeur (Frédéric Maire occupe le poste depuis 2009). Les statuts de la fondation détaillent le mandat de la Cinémathèque suisse qui repose sur les quatre missions suivantes :

- « Recueillir et sauvegarder les archives de la cinématographie, quelle qu'en soit l'origine ;
- Veiller à l'accroissement, à la conservation, à la restauration et à la présentation de ses collections ;
- Constituer un musée national et un centre d'étude de la cinématographie ;
- Servir l'utilité publique et ne viser aucun but lucratif. » (Cinémathèque suisse [sans date])

« Institution nationale de conservation du cinéma et des images en mouvement », la Cinémathèque suisse, installée sur trois sites (Lausanne – Centre administratif et projections, Penthaz – Centre de recherche et d'archivage, et Zurich – Forschungs- und Archivierungs-Zentrum), collectionne en priorité « tous les supports d'information (tous supports confondus Film et Non-Film) qui correspondent aux critères définissant les "Helvetica" dans les domaines

cinématographique et audiovisuel » (Cinémathèque suisse 2015a, p. 12, 16). Cette notion centrale d'« Helvetica » provient de la Loi fédérale sur la Bibliothèque nationale suisse (BNS) qui précise que cette dernière « collectionne les informations imprimées ou conservées sur d'autres supports que le papier, qui : a. paraissent en Suisse ; b. se rapportent à la Suisse, à ses ressortissants ou à ses habitants ou c. sont créés, en partie ou en totalité, par des auteurs suisses ou par des auteurs étrangers liés à la Suisse » (*Loi fédérale sur la Bibliothèque nationale suisse (LBNS ; 423.21)* 1992, art. 3). La BNS coordonne ses activités avec d'autres institutions qui collectionnent, archivent, répertorient et rendent accessibles certaines catégories d'« Helvetica », dont la Phonothèque nationale, la Cinémathèque suisse et les Archives fédérales (*Ordonnance sur la Bibliothèque nationale suisse (OBNS ; 432.211)* 1998, art. 4). La Cinémathèque suisse s'occupe alors particulièrement de la collecte d'informations relatives au « film suisse » (défini dans le cadre de la *Loi fédérale sur la culture et la production cinématographiques (LCin ; 443.1)* 2001, art. 2, 3) et conserve de manière prioritaire les éléments suivants :

- « Films de nationalité suisse (production ou coproduction) ou d'un réalisateur suisse ou domicilié en Suisse.
- Films subventionnés par la Confédération, ou ayant bénéficié d'un financement public ou privé suisse.
- Films comprenant des acteurs, des techniciens (ex : chef opérateur) ou des artistes principaux suisses ou domiciliés en Suisse.
- Version suisse de films étrangers.
- Films tournés en Suisse et documents ou objets se rapportant au domaine cinématographique ou ayant un lien majeur avec la Suisse.
- Objets ou documents cinématographiques suisses documentant l'histoire de la Cinémathèque ou ayant un lien avec les activités de la Cinémathèque. » (Cinémathèque suisse 2015a, p. 13)

La Cinémathèque suisse est reconnue par la Fédération internationale des archives du film (FIAF) comme étant l'une des dix plus importantes cinémathèques du monde grâce à l'étendue, la variété et la qualité de ses collections, qui comptent quelque « 85'000 titres de films, soit 700'000 bobines, des centaines de fonds filmiques, 2,5 millions de photos, 500'000 affiches, 26'000 livres, 720'000 périodiques, 10'000 scénarios, plus de 200 fonds d'archives papier, 240'000 dossiers documentaires, 2000 appareils cinématographiques » (Cinémathèque suisse [sans date]).

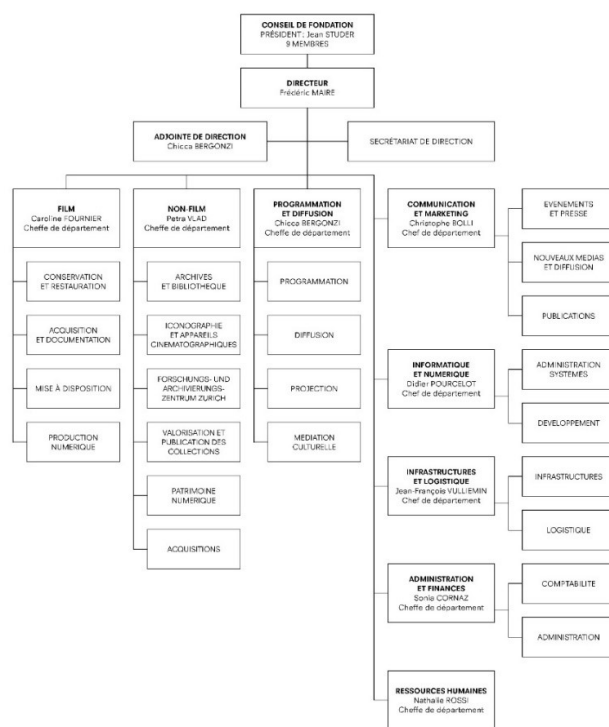
### **1.2.2 Organisation et projets numériques**

En 2010, les secteurs de collection sont regroupés en deux départements distincts : Film et Non-film. Ce dernier collectionne alors tout ce qui a trait au septième art en Suisse, dont les livres, les périodiques, les scénarios, les dossiers documentaires, les photographies, les affiches, le matériel promotionnel mais aussi – ce qui intéresse tout particulièrement la présente étude – les archives institutionnelles et les fonds privés. Récemment restructuré par Petra Vlad (2020), le département compte dorénavant trois secteurs de collection (« Archives et bibliothèque » ; « Iconographie et appareils cinématographiques » ; « Forschungs- und Archivierungs- Zentrum Zürich ») et deux secteurs de services : « Valorisation et publication

des collections » et « Patrimoine numérique » (voir Figure 1 : Organigramme de la Cinémathèque suisse). Jouant un rôle transversal pour accompagner la transition numérique dans la collecte, le traitement, la conservation et la valorisation des collections, le nouveau secteur « Patrimoine numérique », placé sous la responsabilité de Rebecca Rochat, doit répondre principalement aux trois problématiques suivantes : « la numérisation des collections 2D (documents opaques et transparents) et la modélisation 3D ; la mise au point des techniques de récolte des collections nées numérique (acquisition, moissonnage, extraction des données des supports obsolètes) ; l'archivage et la conservation pérenne de ces collections » (Cinémathèque suisse 2020a, p. 33). Ce secteur participe également de manière active au vaste projet initié en 2020, « Plateforme logicielle », qui vise à développer et / ou à acquérir des logiciels métiers pour la gestion, l'archivage et la valorisation des collections analogiques et numériques.

Cette restructuration et ces projets (auxquels s'ajoute la publication en mai 2022 des « Modalités des dépôts numériques à la Cinémathèque suisse » qui précise les formats et les critères d'acceptation de données numériques par l'institution (Cinémathèque suisse 2022a)) témoignent de l'importance grandissante du numérique dans les collections de la Cinémathèque et de la prise de conscience du changement de paradigme que cela représente. Alors qu'en 2009 le projet de construction du Centre de recherche et d'archivage de Penthaz pouvait encore se permettre de ne pas intégrer la conservation et le traitement des données numériques (Cinémathèque suisse 2015b, p. 1), et qu'on peut lire dans la politique de collection publiée en 2015 que « les supports et documents numériques sont encore très peu représentés dans les fonds d'archives » (Cinémathèque suisse 2015a, p. 3), la situation a semble-t-il rapidement évolué.

Figure 1 : Organigramme de la Cinémathèque suisse



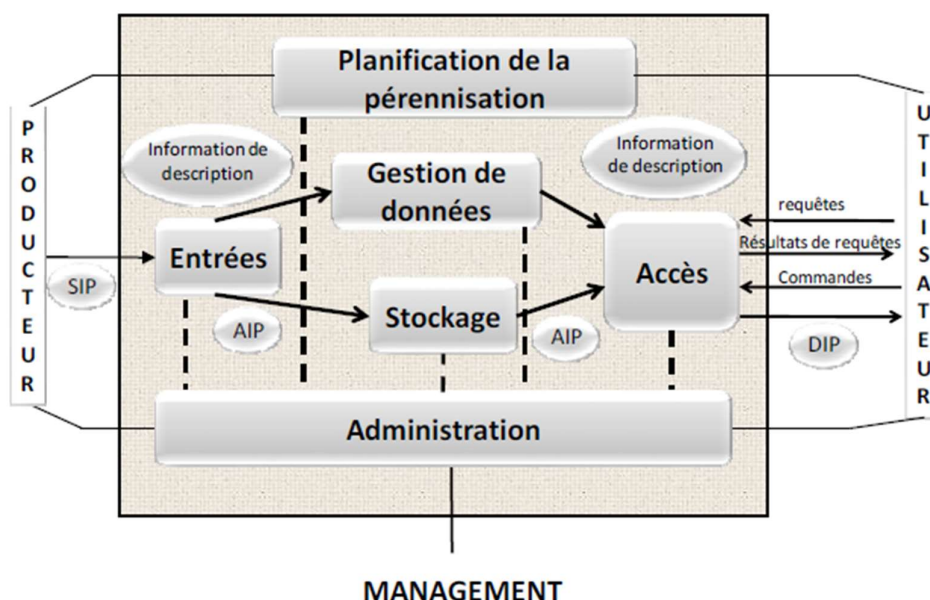
(Cinémathèque suisse 2020b)



### 1.2.3 Archivage des données numériques à la Cinémathèque suisse : le projet *Ingest Manager*

Ingest Manager, logiciel développé par l'équipe informatique de la Cinémathèque suisse (sur des spécifications métiers transmises par les membres du comité technique issus des Départements Film et Non-Film), doit permettre de prendre le virage du numérique pour ce qui est de la collecte et du traitement des données numériques acquises et conservées par la Cinémathèque suisse : que ce soit le matériel promotionnel et de production, les œuvres cinématographiques elles-mêmes mais aussi les supports de données provenant de personnalités ou d'organismes privés qui contiennent des documents numériques répondant à la politique de collection de l'institution (comme des scénarios rédigés par des réalisateurs, des demandes de financement, des photographies de tournage, des échanges de correspondance électronique, etc.). L'archivage de ces fichiers numériques (que ce soit ceux que la Cinémathèque suisse reçoit de la part de tiers, mais également les données numériques qu'elle produit via la numérisation de ses collections analogiques) nécessite des instruments et des procédures *ad hoc*. La Cinémathèque, comme beaucoup d'institutions patrimoniales confrontées aux mêmes problématiques, a choisi d'adopter le modèle *OAIS* (*Open Archival Information System* ou, en français : *Système ouvert d'archivage d'information*), développé initialement dans le domaine de la recherche spatiale (CCSDS) à la fin des années 1990 pour traiter le nombre gigantesque de données produites et reçues. Devenu depuis une norme ISO (ISO 14721 :2012), *OAIS* n'est pas un produit informatique ou une marche à suivre stricte ; c'est « un modèle conceptuel de gestion, de conservation et de préservation à long terme de documents numériques », « un référentiel permettant de saisir toutes les spécificités, les considérations et les acteurs de l'archivage numérique à long terme » (Makhlouf Shabou 2021, p. 1). Le modèle en question définit quatre rôles (« Producteur », « Utilisateur », « Archive » et « Management ») et six entités fonctionnelles (« Entrées », « Stockage », « Accès », « Administration », « Planification de la pérennisation » et « Gestion de données ») :

Figure 2 : Entités fonctionnelles *OAIS*



(Comité Consultatif Pour les Systèmes de Données Spatiales (CCSDS) 2017, p. 4-1)

Le projet de logiciel *Ingest Manager* vise à l'implémentation de *OAIS* au sein de la Cinémathèque suisse. Le modèle *OAIS* définit notamment un *SIP* (*Submission Information Package*, « Paquet d'informations à verser » en français) que l'entité « Entrées » reçoit de la part du « Producteur », paquet qu'il va falloir préparer selon des règles et en suivant un flux de traitement pour générer un *AIP* (*Archival Information Package* ou « Paquet d'informations archivé ») qui sera pris en charge par l'entité « Stockage ». Un important travail a donc été mené à la Cinémathèque pour traduire ce modèle conceptuel en besoins concrets et en tâches spécifiques à réaliser pour garantir l'intelligibilité et l'accessibilité à long terme des données archivées. Parmi ces tâches, on compte notamment l'identification des formats et la conversion vers des formats pérennes, la création de paquets de métadonnées, l'attribution d'identifiants uniques, ou le renommage en masse de fichiers (Cinémathèque suisse 2022b). La mise en production de ce logiciel et son fonctionnement prochain doivent permettre de pallier les insuffisances et défauts de la situation actuelle : beaucoup de tâches sont réalisées manuellement, parfois à double, et de nombreuses vérifications sont nécessaires – autant de facteurs qui engendrent un retard dans le traitement des données.

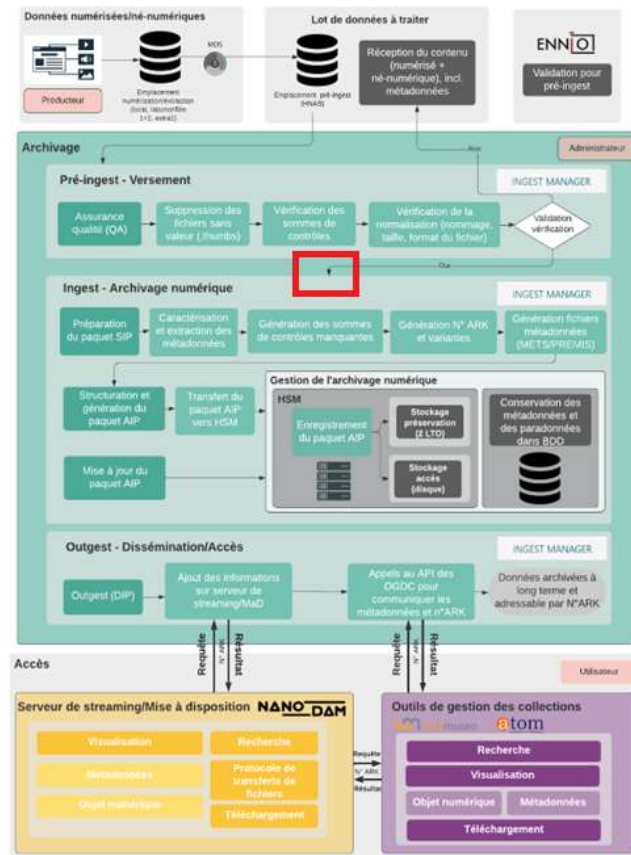
Le flux de travail général ci-dessous (Figure 3 : « Flux numérique, Archivage / INGEST, Non-Film ») a été conçu par le secteur « Patrimoine numérique » de la Cinémathèque : on y voit, au sommet à gauche (dans un carré gris) l'étape d'extraction des données (c'est-à-dire les données provenant des logiciels de numérisation ou via la réalisation d'une image disque « forensique » d'un support de données, effectuée grâce au logiciel *Aaru*<sup>1</sup>) ; puis, au sein du grand rectangle vert qui symbolise l'archivage, un premier étage intitulé « Pré-Ingest – Versement » qui décrit les premières étapes à suivre après l'acquisition du matériel ou l'extraction des données ; et enfin un second étage au sein duquel sont décrites les actions à réaliser pour l'« Ingest », c'est-à-dire pour l'archivage numérique proprement dit (préparation du *SIP* et génération du *AIP* pour un archivage pérenne dans le *Hardware Security Module* – *HSM* – de la Cinémathèque suisse – autrement dit : un stockage de haute sécurité sur des bandes *LTO*).

L'évaluation et le tri archivistiques (symbolisés dans le diagramme par un rectangle rouge, ajouté par nos soins), sujet de ce Mémoire de Master, trouveraient donc leur place au sein de ce flux de travail entre la partie « Pré-Ingest – Versement » (puisque les premières étapes de curation de données ont déjà été réalisées et qu'un contrôle de qualité, via des sommes de contrôle, a été entrepris pour s'assurer de l'intégrité des données) et la section « Ingest – Archivage numérique » (puisque'il s'agit alors de préparer les données que l'on souhaite archiver dans le *HSM*, c'est-à-dire les seules données « sélectionnées », celles que la Cinémathèque a estimé nécessaire de conserver). Les étapes d'archivage proprement dit ne portent effectivement « que » sur les données à préserver.

---

<sup>1</sup> Les détails techniques et méthodologiques de l'extraction de données donnera lieu à une présentation de la part de Robin François, archiviste numérique, et Rebecca Rochat, en septembre 2022 à Glasgow lors de la conférence *iPres 2022: The 18th International Conference on Digital Preservation*. (François, Rochat 2022).

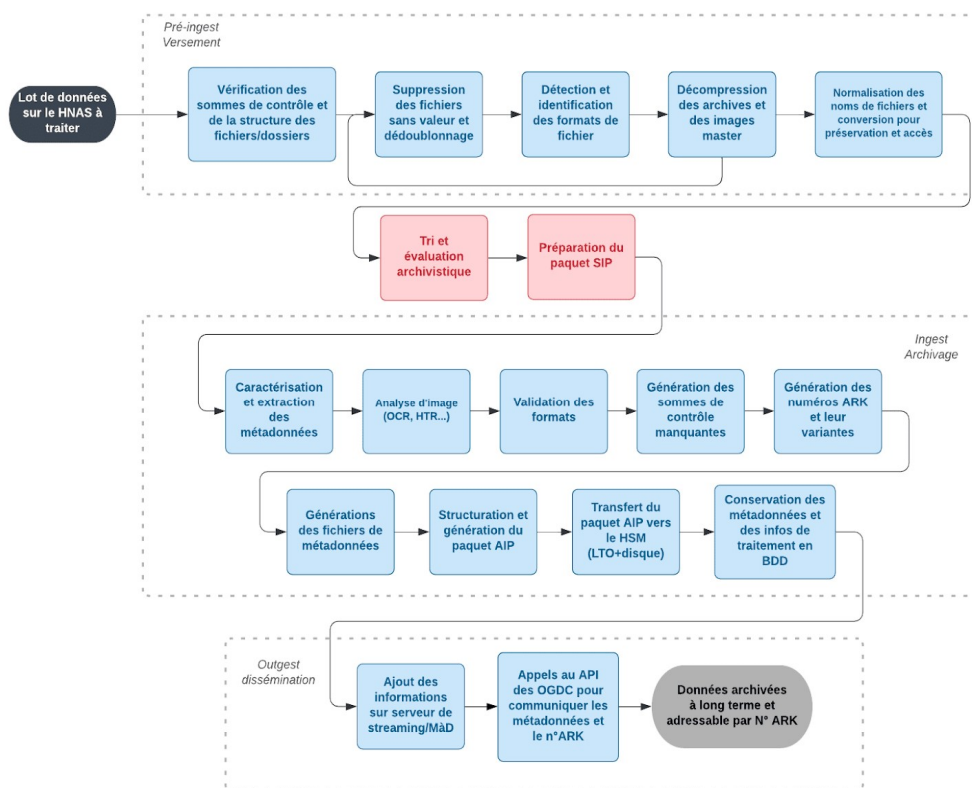
Figure 3 : « Flux numérique, Archivage / INGEST, Non-Film »



(Rochat 2021)

Un flux décrivant les traitements imposés aux lots de données au Département Non-Film de la Cinémathèque suisse précise un peu plus la place et le contexte de ce Mémoire de Master puisque le diagramme ci-dessous (Figure 4 : « Flux de traitements des lots de données au Département Non-Film ») ne prend plus en considération tout le cycle de l'archivage (le premier graphique intégrait la mise à disposition des contenus via un serveur de streaming, appelé *NanoDam* par la Cinémathèque), mais uniquement les étapes de « Pré-Ingest » et d'« Ingest » proprement dites – on retrouve ainsi la case « Tri et évaluation archivistiques » après l'extraction et la curation des données, mais avant la soumission du paquet à archiver (*SIP*).

Figure 4 : « Flux de traitements des lots de données au Département Non-Film »



(François 2021)

Tout l'enjeu de ce Mémoire de Master consiste donc à affiner la tâche jusque-là trop généraliste « Tri et évaluation archivistiques » en distinguant les étapes nécessaires permettant de passer des données « brutes », telles que récupérées auprès d'un tiers après l'extraction, aux données « sélectionnées », soit celles que l'on désire à terme archiver dans le *HSM*.

## 1.3 Revue de littérature

### 1.3.1 Projets précédents et méthodologies existantes

Chronologiquement, le premier projet d'envergure qui vaut la peine d'être cité est celui mené conjointement par les universités de Manchester et d'Oxford entre 2005 et 2007 et intitulé *Paradigm* (Paradigm Project 2007). Acronyme de **Personal Archives Accessible in Digital Media**, cette recherche, qui a débouché sur la publication d'un manuel (*Workbook*) contenant des recommandations d'actions et d'outils ainsi que des tests pratiques, désirent combler une lacune : tous les projets menés jusque-là à propos de la préservation et l'accessibilité des documents numériques à long terme concernaient les archives d'État ou d'entreprises, mais il n'existait quasiment rien à propos des archives privées. *Paradigm* étudie donc le traitement des archives personnelles nées numérique en prenant l'exemple des documents de personnalités politiques et en suivant un cycle de vie fait de sept grandes étapes, des contacts avec les producteurs d'archives à leur mise à disposition des chercheurs, en passant par leur stockage dans des dépôts spécialisés. Un chapitre est donc consacré spécifiquement à l'évaluation. Les approches recommandées par *Paradigm* sont notamment les suivantes : une évaluation fonctionnelle des documents (soit une approche systématique et de haut niveau

qui analyse les fonctions et activités du créateur des archives et leur contexte de production plus que le contenu individuel des documents) et une stratégie ascendante (*bottom-up*) qui vise à analyser un échantillon de dossiers pour contrôler la fiabilité des intitulés au niveau d'une série et ne pas traiter les archives par documents ou dossiers individuels. Enfin, parmi les solutions pratiques proposées figurent : l'encouragement des meilleures pratiques dès la création des documents ; l'obtention d'informations contextuelles auprès des donateurs ; l'élimination des doublons et l'élimination des fichiers liés au système d'exploitation et aux applications. Aux côtés de ces recommandations pratiques, *Paradigm* reconnaît qu'il est difficile d'établir des « critères d'évaluation » généraux : « Aside from straightforward rules such as removing duplicates, and material that has no long-term research potential, each professional group will have its own concerns » (Paradigm Project 2007, p. 44).

Deux ans plus tard, en 2009, débute un nouveau projet, intitulé AIMS (pour « AIMS Born Digital Collections: **A**n Inter-Institutional **M**odel for **S**tewardship ») qui reprend une partie des objectifs de *Paradigm* (il reproche cependant à ce dernier de s'être trop concentré sur l'acquisition des collections) et qui vise à fournir une approche unifiée de la prise en charge des documents nés numérique. Suite aux études de cas menées dans les institutions partenaires (des bibliothèques et universités américaines principalement), des objectifs principaux ont été définis (décomposés ensuite en résultats attendus, points de décision et tâches spécifiques à réaliser) pour les quatre grandes fonctions archivistiques identifiées : « Collection Development », « Accessioning », « Arrangement and Description » et « Discovery and Access ». Comme on peut le constater, AIMS ne retient pas l'évaluation comme une fonction à part entière et s'en explique ainsi : « Appraisal is also not defined as a specific, separate function. Rather, appraisal activities are included in any or all of the first three functions within this framework » (AIMS Work Group 2012, p. 2). S'il n'est guère possible de résumer ici tous les objectifs par fonction, citons seulement les tâches suivantes, qui nous paraissent pertinentes pour l'évaluation. Pour la fonction « Accessioning » : créer un inventaire au moment du transfert des documents, récupérer les métadonnées des fichiers et du système de fichiers, documenter les restrictions potentielles ou le matériel qui a été accidentellement acquis, identifier les actifs en double. Pour la fonction « Arrangement and Description » : évaluer les documents numériques et leurs relations avec les autres documents et les acquisitions précédentes<sup>2</sup>, évaluer l'authenticité et l'intégrité des documents, déterminer ou proposer un classement, déterminer un niveau de description, identifier les documents à conserver, à éliminer et ceux dont il faut restreindre l'accès<sup>3</sup>. Outre ces tâches archivistiques, AIMS propose quelques tests de logiciels (dont *Karen's Directory Printer*) et, plus intéressant encore, il établit une liste des exigences fonctionnelles d'un outil à développer (« Appendix H : Technical Development »). Deux conclusions du projet AIMS doivent enfin être relevées, car elles mettent en perspective les difficultés auxquelles les institutions sont confrontées. L'une à propos des outils à disposition, l'autre concernant les étapes à suivre : « While several tools fulfilled some required needs, no single, open-source solution was identified for arrangement and description » (AIMS Work Group 2012, p. VI) ; « This is a practical approach but also a recognition that there is no single solution for many of the issues that institutions face when dealing with born-digital collections » (AIMS Work Group 2012, p. VIII).

---

<sup>2</sup> « Assess the born-digital records and relationships with other material and previous accessions (if they exist) » (AIMS Work Group 2012, p. 37)

<sup>3</sup> « Identify records for retention, restriction, and removal (if possible); identify relevant levels of access and edit metadata accordingly. » (AIMS Work Group 2012, p. 39)

Deux projets plus récents, qui ont donné lieu à des *workflow*, méritent encore d'être mentionnés. Le premier s'intitule *OSSArcFlow Project* et a été mené entre 2017 et 2020 par l'institut EduCopia et visait à étudier la manière dont les logiciels *Open Source* (OSS, *Open Source Software*) pouvaient soutenir la curation numérique et l'archivage à long terme. L'équipe de projet a notamment étudié les flux de travail de douze institutions partenaires utilisant *BitCurator*, *ArchivesSpace* et *Archivematica*. Plusieurs étapes sont communes à quasiment toutes les institutions : création d'un dossier d'acquisition, création d'une image disque des supports physiques, saisie de métadonnées descriptives et techniques, conservation dans des environnements de préservation dédiés et dans des dépôts, et description (à des niveaux variables) des documents pour permettre la découverte et l'accès (instruments de recherche notamment). En revanche, certaines étapes ne sont pas partagées aussi unanimement par toutes les institutions comme la mise en quarantaine des fichiers, les contrôles anti-virus, la documentation de la structure des fichiers sur les disques, l'attribution d'identifiants uniques, le renommage suivant des conventions institutionnelles, ou encore la déduplication des actifs redondants (Post et al. 2019). Le projet a finalement accouché d'une méthodologie unifiée comprenant treize étapes pour l'archivage des données numériques : *Common Steps in OSS Born-Digital Archival Workflows*, comprenant également des recommandations d'outils. Parmi les étapes listées, citons : « Create disk image », « File identification & format characterization », « Check file integrity & ensure fixity », « Create accession record », « Analyze and identify sensitive content », « Analyze forensic/technical metadata », « Create/extract digital object metadata » ou encore « Assemble AIP » (Chassanoff, Post 2020, p. 10-29).

Le second est un projet suisse, réalisé par le KOST-CECO, soit le « Centre de coordination pour l'archivage à long terme de documents électroniques », dont les résultats viennent seulement d'être publiés (juin 2022) : « 20-039 Collections de fichiers ». Ce projet vise à étudier la manière dont doivent être pris en charge les « collections de fichiers », c'est-à-dire les documents, qu'ils proviennent d'une administration ou de personnes privées, qui ne sont pas inscrits dans un système de gestion *ad hoc* (type *Gestion Électronique des Documents*) ou une base de données par exemple. Dans le monde germanophone, un tel sujet avait déjà donné lieu à un atelier à Munich en 2016 intitulé *Kreative digitale Ablagen und die Archive*, où la collection de fichiers était définie de la manière suivante :

« Menge von Einzeldateien, die von einem oder mehreren Bearbeitern zur Erledigung einer oder mehrerer Aufgaben über einen bestimmten Zeitraum erstellt und nach individuellen Ordnungskriterien zusammengestellt wurden. Die Dateien liegen auf einer Ebene und/oder hierarchisch in einer Verzeichnisstruktur vor. Es können in einer Dateisammlung unterschiedlichste Dateiformate enthalten sein. » (Miegel, Schieber, Schmidt 2017, p. 7)

Ces « collections de fichiers » correspondent bien au type de documents et de répertoires auxquels est confrontée la Cinémathèque suisse avec les supports de données provenant de personnes ou d'organismes privés. Le KOST-CECO a proposé un *workflow* pour leur traitement qui contient six étapes, dont l'une nous paraît particulièrement intéressante pour notre sujet d'étude : « Analyse », décomposée en trois phases, « Technische Analyse », « Archivfachliche Analyse » et « Abschluss Analyse ». L'analyse technique comprend notamment la récupération des métadonnées descriptives et techniques, l'inventaire des formats de fichiers, l'établissement d'une liste de doublons, le repérage des fichiers système, des fichiers temporaires et des dossiers compressés. L'analyse « archivistique » comprend l'étude du classement (*Ordnungsstruktur*) et de la volumétrie (*Mengengerüst*) du répertoire

afin de mener une évaluation *top-down* des dossiers (« *die Ablieferung top-down zu bewerten* ») de sorte à identifier les éléments qui ne méritent pas d'être archivés (cette étape comporte également l'identification des données sensibles) (KOST-CECO 2022a).

### 1.3.2 Listes d'outils

Plusieurs communautés, associations ou individus, au niveau national ou international, se sont attelés à la tenue d'une liste d'outils potentiellement intéressants pour la prise en charge d'archives numériques. Ces listes couvrent des domaines distincts et adoptent des formes différentes.

Le *Service Interministériel des Archives de France (SIAF)* propose par exemple, dans son carnet numérique intitulé « Modernisation et archives », la « boîte à outils numériques de l'archiviste » sous la forme d'une carte mentale (*mind map*) regroupant des outils servant non seulement à la gestion de projets ou à la veille technologique mais aussi au traitement de fichiers numériques, dans les domaines suivants : « Édition de fichiers », « Renommer », « Nettoyer des données ou des fichiers, visualiser, comparer », « Gestion des formats et de la pérennisation ». (Coline1 2019). Un même type de carte mentale, plus ciblée – la « Carte des outils utiles pour l'archivage électronique » –, a été conçue par Julien Benedetti (archiviste aux Archives départementales des Bouches-du-Rhône et membre du Conseil d'administration de l'Association des archivistes français) à partir des échanges professionnels inter-institutions qui se tiennent sur le fil twitter de l'association (Benedetti 2021). Sont actuellement référencés les outils de « pré-versement » répartis dans trois catégories : « Pérennité », « Intégrité » et « Classement et tri ». La dernière section, la plus intéressante pour notre propos, contient quatre sous-catégories : « Explorer une arborescence » (*Archifiltre*, *WinDirStat* et *Pir* sont mentionnés), « Renommage des fichiers », « Dédoublonnage » (avec entre autres *AllDup* et *WinMerge*) et « Réorganiser un fonds » (*Octave*, *Docuteam Packer*).

À ces listes établies par des organismes français répondent également des plateformes allemande et suisse. *Nestor*, mot-valise composé de « network » et « storage » est le réseau compétent pour la préservation et l'accessibilité à long terme des documents numériques en Allemagne. Hébergé par la Bibliothèque nationale allemande, *Nestor* tient à jour une liste d'une centaine d'outils répartis selon leur domaine d'expertise dont « Übertragung », « Formaterkennung », « Formatmigration », « Packen », « Metadaten » ou « Nutzung ». Le domaine qui nous intéresse particulièrement (« Bewertung » – Évaluation) comporte presque trente outils (dont *TreeSize*, *AllDup*, *WinMerge*, etc.) (Nestor - Deutsche Nationalbibliothek 2022). L'équivalent suisse, c'est le réseau KOST-CECO : on trouve sur leur site internet quelques outils développés à l'interne (comme *KOST-Val* pour la validation des formats, ou *KOST-Simy* pour la comparaison visuelle automatisée d'images) ainsi qu'une liste d'une quarantaine d'outils établie lors d'un projet de *Workflow* (« KOST-Projekt 20-039 Dateiablage ») dont il a été question ci-dessus. Les outils doivent servir à l'accomplissement de tâches lors des étapes de traitement suivantes : « Sichtung », « Uebernahme », « Analyse », « Aufbereitung » et « Erschliessung » (KOST-CECO 2022b).

Au niveau international enfin, il existe une plateforme gérée par la *Community Owned digital Preservation Tool Registry (COPTR)* qui compte parmi ses partenaires les organismes suivants : *The Digital Curation Centre (DCC)*, *The Digital Preservation Coalition (DPC)*, *National Digital Stewardship Alliance (NDSA)*, le réseau *Nestor* ainsi que *The Open Preservation Foundation (OPF)*. Le projet *COPTR* visait précisément à regrouper en un seul lieu (un registre wiki) toutes les listes d'outils tenues à jour par différents organismes qui se

chevauchaient souvent, sans se recouper entièrement. Pensé comme un registre permettant aux archivistes de trouver facilement le logiciel dont ils ont besoin, *COPTR* compte actuellement 568 outils (et 21 *workflows* partagés). Les outils sont classés selon leurs fonctionnalités (une cinquantaine dont « Appraisal » – 11 outils, « De-Duplication » – 18 outils, ou « File Management » – 37 outils) mais aussi selon le cycle de vie des documents numériques, et le type de fichiers et les formats qu'ils prennent en charge (Community Owned digital Preservation Tool Registry (COPTR) 2021).

## 1.4 Objectifs et questions de recherche

La revue de littérature ci-dessus et le contexte institutionnel dans lequel s'inscrit ce Mémoire de Master nous permettent de préciser les objectifs de ce travail et de formaliser les questions de recherche auxquelles nous essaierons de répondre. Il s'agira avant tout de combler deux lacunes principales dans la recherche actuelle : un manque d'informations concernant les outils permettant d'effectuer un tri archivistique et une absence de détails pratiques quant aux tâches spécifiques à mettre en œuvre pour mener une évaluation.

Les listes existantes d'outils, aussi détaillées et exhaustives soient-elles, ne précisent pas de quelle manière les logiciels fonctionnent réellement : on peut certes connaître le nom des outils qui réalisent une tâche et obtenir un bref descriptif sur leur champ d'application, mais aucune liste ne fait état de leurs fonctionnalités précises, ce qui rend les comparaisons, et donc le choix de l'outil, très difficiles. Rappelons par exemple que la plateforme *COPRT* ne propose pas moins de 18 outils qui permettent de réaliser un dédoublement des éléments redondants, mais quel logiciel doit choisir une institution patrimoniale comme la Cinémathèque suisse ? En outre, il arrive fréquemment que les projets que nous avons mentionnés conseillent des outils pour réaliser les étapes de leur flux de travail, mais là aussi : les détails manquent. Enfin, les études de cas consultées ((Kim, Dong, Durden 2006 ; Forstrom 2009 ; Wilsey et al. 2013 ; Oestreicher 2013 ; Shein 2014 ; Meister, Chassanoff 2014 ; Sloyan 2016 ; Vinh-Doyle 2017 ; Belovari 2017 ; Schneider et al. 2019) se limitent soit au test d'un outil particulier, soit elles passent sous silence les raisons qui ont motivé leur choix (et elles ne précisent pas toujours les logiciels qui ont été testés avant d'arrêter leur décision). Une **première partie** de ce Mémoire de Master (« 3. Analyse des fonctionnalités logicielles ») tentera donc de combler cette lacune en répondant aux questions suivantes :

- Quelles sont les fonctionnalités *détaillées* offertes par les outils informatiques disponibles sur le marché permettant de mener à bien des tâches d'évaluation et de tri archivistiques ?
- Quelles sont les fonctionnalités les plus *importantes* pour réaliser ladite tâche et quels sont les outils qui proposent ces fonctionnalités ?
- Quel outil, ou quelle combinaison d'outils, la Cinémathèque suisse devrait adopter et implémenter pour le traitement de ses collections numériques ?

Les méthodologies précédemment publiées (de *Paradigm* en 2005-2007 au KOST-CECO en juin 2022) souffrent des mêmes défauts. Premièrement, elles cherchent la plupart du temps à couvrir l'ensemble du cycle de vie des données numériques, de leur accession auprès d'un donateur, à leur mise à disposition du grand public. Si certaines se montrent plus précises sur les tâches d'évaluation à accomplir (c'est le cas du KOST-CECO qui entre dans les détails de l'« Analyse » en évoquant l'étude de la structure de l'arborescence et du plan de classement initial pour mener une évaluation *top-down*), d'autres ne font qu'évoquer l'évaluation, voire l'intègrent aux fonctions archivistiques voisines comme la description ou la classification



(AIMS). Deuxièmement, et c'est une conséquence du périmètre très large qu'ils adoptent, ces *workflow* ne sont pas assez détaillés pour permettre leur mise en pratique immédiate par une institution patrimoniale. Si la plupart d'entre eux mettent en avant la difficulté de formaliser et de généraliser des tâches et des critères qui dépendront fortement du contexte et de la politique de l'institution, ces flux de travail n'en demeurent pas moins trop généraux pour s'appliquer à une situation réelle. Ainsi, à titre d'exemple, alors que le dédoublonnage est souvent évoqué, aucune méthodologie n'explique véritablement et dans le détail comment procéder. Or, la distance est immense entre dire théoriquement qu'il est nécessaire de dédoubler et se trouver réellement confronté à plusieurs milliers de fichiers redondants dans un fonds d'archives numériques. La **seconde partie** de ce Mémoire de Master (« 4. Méthodologie de tri archivistique ») tentera donc de proposer une méthodologie de tri et d'évaluation qui détaille des tâches et des sous-tâches, et qui mette aussi en avant les points d'analyse / de décision, les traitements possibles, et les outils informatiques avec lesquels travailler. Il s'agira alors de répondre aux questions suivantes :

- Quelles sont les tâches et les sous-tâches (plus détaillées) à mettre en œuvre pour réaliser une évaluation et un tri archivistiques de données numériques ?
- Quels sont les points d'analyse à prendre en compte avant de décider du sort final d'un élément ?
- Quels sont les instruments, les méthodes et les outils à utiliser pour mener à bien les tâches évoquées ?

## 2. Méthodologie générale

### 2.1 Typologie de la recherche

Pour délimiter le périmètre général de cette étude, il convient de la situer parmi les types de recherche possibles en Sciences de l'Information. Rappelons donc avant tout qu'il s'agit d'une recherche dite « appliquée » puisqu'elle « consiste à trouver des solutions à des problèmes pratiques et que la connaissance peut être immédiatement axée sur l'action ou la prise de décision » (Fortin 2016, p. 19) : il s'agit effectivement de mener une enquête centrée sur les besoins des archivistes, confrontés à de gigantesques masses de données, et de (leur) proposer des outils et des tâches spécifiques pour réaliser l'évaluation et le tri archivistiques des supports de données. Comme on l'a constaté en parcourant la littérature secondaire, ce domaine d'étude est encore relativement peu développé, de manière aussi précise et pratique tout du moins, et nous désirons donc mener une recherche de type « exploratoire », en ayant recours à l'étude de cas comme approche générale et moyen de collecter des données. Si l'étude de cas peut notamment être définie comme « une approche de recherche empirique qui consiste à enquêter sur un phénomène, un événement, un groupe ou un ensemble d'individus, sélectionné de façon non aléatoire, afin d'en tirer une description précise et une interprétation qui dépasse ses bornes » (Roy 2010, p. 206-207), nous retiendrons de cette définition tirée de la recherche en sciences sociales les éléments suivants : l'aspect empirique de la démarche ; la sélection non aléatoire de l'échantillon ; l'étude détaillée d'un phénomène ou d'une situation ; et surtout l'interprétation qui « dépasse ses bornes ». Car c'est bien au moyen de tests en situation réelle, sur des fonds d'archives numériques précis et grâce à des outils sélectionnés selon quatre critères (voir « 3.1.2 Critères de sélection des outils testés »), que nous espérons proposer une méthodologie générale et applicable à d'autres ressources numériques conservées par la Cinémathèque.

### 2.2 Échantillon

L'échantillonnage par choix raisonné « demeure la stratégie privilégiée de l'étude de cas » (Fortin 2016, p. 198) et nous ne dérogerons pas à cette règle : la Cinémathèque suisse a ainsi sélectionné deux fonds d'archives numériques qui pourraient se prêter à la comparaison des solutions logicielles et à l'établissement d'une méthodologie de tri archivistique : le Fonds Francis Reusser et le Fonds Ana Simon.

Dans le premier cas, il s'agit d'un disque dur externe ayant appartenu au réalisateur vaudois Francis Reusser. Cinéaste militant, politiquement très engagé dès la fin des années 1960 dans les combats sociaux et culturels de son temps, Francis Reusser naît à Vevey en 1942. Après une formation de photographe, Reusser entame un apprentissage de caméraman à la Télévision Suisse Romande. Sa carrière de cinéaste débute en 1964 avec *Antoine et Cléopâtre*, suivi notamment par son premier long métrage de fiction, *Vive la mort*, « film plein de rage et d'humour, rebelle contre la société et les pères » (Maire 2020), présenté à Cannes en 1969. L'année suivante, on lui doit l'un des premiers films tournés dans les camps palestiniens, avec le collectif Rupture, le documentaire *Biladi, une révolution*, puis avec Patricia Moraz au scénario, il tourne *Le Grand Soir* en 1976, « réflexion désabusée sur ce fameux "grand soir" qui n'est pas vraiment arrivé en 1968 » (Maire 2020). Dans les années 1980-1990 suivront des films plus poétiques (*Seuls*, 1981), ou adaptés entre autres de l'œuvre de Charles-Ferdinand Ramuz (*Derborence*, 1985 ; *La Guerre dans le Haut-Pays*, 1999), avant de s'atteler bien des années plus tard (2012) à une adaptation de Jean-Jacques Rousseau :

*Ma Nouvelle Héloïse*. Francis Reusser est également actif au sein du *Ciné-atelier*, une société de production fondée en 1998 avec Emmanuelle de Riedmatten, réalisatrice du documentaire *Carole Roussopoulos, une femme à la caméra* en 2011. Francis Reusser tourne encore *La Séparation des traces* en 2018, avant de s'éteindre, à Bex, en avril 2020. Ses archives ont rejoint la Cinémathèque suisse après son décès dans le courant de l'année 2020, par l'intermédiaire de son fils, Jean Reusser (Neeser 2022). Elles contiennent entre autres des documents administratifs, de la comptabilité, des dossiers de production, des photos, des articles écrits par Francis Reusser, une revue de presse et le disque dur externe, objet de l'étude de cas à venir.

Le second lots d'archives numériques à notre disposition pour réaliser les études de cas est constitué d'une dizaine de clés USB (Figure 5 : Quelques-unes des clés USB d'Ana Simon, Fonds Simon) ayant appartenu à la réalisatrice, metteuse en scène, poétesse et écrivaine Ana Simon (1938-2018). Épouse de François Simon et belle-fille de Michel Simon, Ana Simon leur a consacré deux documentaires : *François Simon, la présence* (1986) et *Simon, père et fils* (1995). Née en 1938 en Roumanie, Ana Simon est également l'autrice de plusieurs recueils de poèmes dont *Vivre* (1981), *Entrevision* (1985), *Jardin désolé* (1995) ou *Les Muses endormies* (2004) – ce dernier est d'ailleurs illustré par son amie, la peintre allemande, Margarethe Krieger, dont le nom apparaît à plusieurs reprises dans les documents stockés sur les clés USB. Polyglotte, maîtrisant parfaitement le français et l'espagnol et ayant étudié la littérature comparée, Ana Simon fut aussi traductrice, notamment de Mircea Eliade, Marin Sorescu et Miguel de Unamuno. L'une de ses dernières réalisations, « parcours entre sa poésie et sa ville d'adoption » (Maire 2019), fut *Sortilèges de Genève*. Les Papiers Ana Simon ont été rassemblés à la Cinémathèque entre 2013 et 2019 et comptent peu de documents papiers : 2 boîtes d'archives, 2 cartons de déménagement et 1 carton plat. Il s'agit alors principalement d'éléments épars de correspondances, de documentation, de traductions, de scénarios et de projets de film ainsi que de quelques photographies (sont notamment documentés les films suivants : *La Petite Vendeuse de lampes*, 1987 et *François Simon, la présence*, 1986). Un Fonds Ana Simon est également conservé à Barcelone (Biblioteca d'Humanitats Universitat Autònoma) et à la Bibliothèque de Genève (Tourn 2019).

Figure 5 : Quelques-unes des clés USB d'Ana Simon, Fonds Simon



Si pour le premier de ces deux fonds d'archives numériques (Francis Reusser), une première analyse avait été réalisée par la Cinémathèque (par le biais du logiciel français d'étude des arborescences, *Archifiltre*) permettant de connaître approximativement de quoi était fait le disque dur externe (2'455 dossiers, 37'589 fichiers, pour un poids total de 240 Go, et une arborescence se développant sur 14 niveaux de profondeur), le second lot (les clés USB d'Ana Simon) n'avait encore donné lieu à aucune analyse préliminaire.

Ces échantillons présentent les avantages suivants qui justifient leur sélection : il s'agit de deux types de supports distincts (disque externe *versus* clés USB) qui rendent donc possible une comparaison (existe-t-il une méthodologie distincte suivant les supports ? à quoi faudrait-il être alors particulièrement vigilant ?) ; une double étude de cas permet également d'atténuer les biais en confrontant les hypothèses ; enfin, l'un au moins de ces deux échantillons contient suffisamment de dossiers et de fichiers, de formats divers et enregistrés à des niveaux de profondeur différents, pour qu'une approche *macroscopique* qui mette à profit les outils actuellement disponibles en termes de traitement de masse et d'automatisation se justifie.

## 2.3 Périmètre de la recherche

Cette recherche portera sur les supports de données et leur contenu général, c'est-à-dire sur les dossiers qui composent l'arborescence des répertoires et sur les fichiers que ces dossiers contiennent. Nous prendrons principalement connaissance du contenu des supports via l'extraction de métadonnées (définies de la manière suivante : « Information that characterizes another information resource, especially for purposes of documenting, describing, preserving or managing that resource » (InterPARES 2018)) qui sont associées aux dossiers et fichiers, ainsi que par le traitement et l'analyse de ces mêmes métadonnées dans des logiciels *ad hoc*. Comme annoncé dans l'introduction de cette étude, il n'est effectivement pas question d'analyser le contenu individuel de chaque élément enregistré sur le support de données, mais bien d'étudier quelles sont les possibilités de traitement *en masse*, via l'analyse des métadonnées uniquement. Les développements actuels dans le domaine de l'intelligence artificielle ou du *data mining* permettent certes d'analyser de grandes quantités de données, et donc le contenu des fichiers, mais ces algorithmes nécessitent des compétences informatiques poussées que nous ne possédons pas, et ce champ de recherche devrait faire l'objet d'une étude séparée. La comparaison des images et des fichiers que nous abordons à la fin de la partie sur les outils et dans la méthodologie générale sera donc une exception, et une timide incursion dans le domaine de l'étude des contenus.

D'autre part, notre méthodologie ne cherchera pas à fixer des critères intangibles pour l'élimination ou la conservation des données. Si l'établissement de critères appartient pleinement au domaine de l'évaluation archivistique, dans le but de déterminer la valeur des documents (que l'on pense aux critères formulés par Boles et Young et articulés en trois modules distincts : valeur de l'information, coûts de conservation et conséquences des décisions résultant de l'évaluation (Young, Boles 1991) ou à la mesure de la qualité des archives issues d'une évaluation (Makhlouf Shabou 2011)), les archives privées de créateur résistent quelque peu à cette tentative de formalisation. En outre, la politique de collection de la Cinémathèque suisse est en cours de refonte, pour intégrer pleinement les documents numériques dans son champ d'action, et les critères d'acquisition de la Cinémathèque suisse que nous mentionnions précédemment (« 1.2.1 Histoire et mandat de collection ») ne peuvent être « opérationnalisés » pour produire des audits sur les contenus et proposer des

recommandations automatiques d'actions via un logiciel informatique. Nous proposons donc surtout une méthodologie pour l'identification des documents et la prise de connaissance du contenu des supports de données ainsi que des points d'attention et des instruments d'analyse pour faciliter la prise de décision.

Enfin, les deux fonds présentés ci-dessus (« 2.2 Échantillon ») seront principalement utilisés pour tester les logiciels sélectionnés et découvrir quelles sont les micro-fonctionnalités qu'ils proposent pour réaliser une tâche. Ces fonds nous serviront également pour les études de cas de la seconde partie, afin d'identifier les tâches nécessaires à mener. En revanche, nous ne procéderons pas à une étude de cas exhaustive et linéaire, visant à « traiter » réellement et complètement, en suivant le *workflow* proposé de A à Z, ces deux fonds d'archives, de manière à « quantifier » le nombre d'éléments soumis à l'élimination (partielle ou totale) ou jugés de valeur suffisamment élevée pour être archivés. Méthodologiquement, cela n'aurait guère de sens puisque le flux de travail proposé doit pouvoir s'appliquer à d'autres supports de données que ceux de Francis Reusser et Ana Simon ; techniquement, nous ne possédons pas les droits d'écriture sur les serveurs de la Cinémathèque qui nous auraient permis d'éliminer « réellement » les fichiers sans valeur archivistique ; enfin, le temps nécessaire à l'évaluation d'un fonds aussi volumineux que celui de Francis Reusser demande une connaissance approfondie de l'œuvre du cinéaste et de ses archives, que nous ne pouvons acquérir dans un temps aussi restreint. Les deux fonds en question serviront ainsi d'exemples pour les analyses et les tâches proposées.

## 3. Analyse des fonctionnalités logicielles

### 3.1 Méthodologie

#### 3.1.1 Critères de sélection des fonctionnalités

Pour déterminer quelles étaient les fonctionnalités informatiques à évaluer en priorité, nous nous sommes basés sur différentes sources. En premier lieu, il s'agissait de répondre aux attentes et aux objectifs formulés par le mandataire de ce Mémoire de Master, en l'occurrence la Cinémathèque suisse et plus particulièrement le Département Non-Film et son secteur « Patrimoine numérique ». Le cahier des charges de cette étude mentionnait alors explicitement les tâches suivantes : « prendre connaissance de l'arborescence », « réaliser des audits sur le contenu », « réaliser le dédoublonnage des fichiers », « identifier les doublons visuels », « gérer le versioning des fichiers », « évaluer le contenu des fichiers » ; autant d'actions devant permettre d'esquisser une méthodologie pour le tri archivistique de supports de données. Parmi les tâches listées, nous avons retenu celles qui nous paraissaient les plus abordables d'un point de vue technique : comme nous l'avons expliqué précédemment, en fixant le périmètre de cette recherche, la prise de connaissance et l'évaluation des *contenus* des documents, quand ces derniers sont extrêmement nombreux, nécessitent des compétences poussées en informatique (pour l'extraction et la préparation des données – domaine connu sous le nom de *data curation* – afin de mener des études de *data mining*), ou en intelligence artificielle si on désire par exemple utiliser des algorithmes d'auto-classification ou le traitement automatique du langage naturel (*Natural Language Processing*, *NLP*). Nous nous sommes donc concentrés sur les tâches qui pouvaient être réalisées via des logiciels actuellement disponibles sur le marché en adoptant une démarche centrée sur l'utilisateur : il s'agissait de se mettre à la place d'un archiviste confronté à un support de données contenant plusieurs (dizaines de) milliers de fichiers qu'il devait traiter dans un temps relativement restreint – sans évaluer le contenu de chaque dossier ou fichier individuellement.

En outre, l'étude des précédentes méthodologies – établies notamment par de grandes institutions américaines et présentées dans la revue de littérature – a confirmé la pertinence d'analyser en priorité ces fonctionnalités informatiques : l'extraction des métadonnées, l'analyse du contenu des supports (arborescence et volumétrie) ou le dédoublonnage sont des tâches fréquemment mentionnées dans les *workflows* existants, accompagnés le plus souvent d'une liste d'outils permettant de les réaliser.

Nous avons retenu en tout quatre fonctionnalités principales : « Extraction de métadonnées / Récolement »<sup>4</sup> ; « Analyse de l'arborescence et de la volumétrie »<sup>5</sup> ; « Recherche de

---

<sup>4</sup> Le terme « extraction de métadonnées » fait référence à la fonctionnalité informatique qui consiste à extraire et compiler dans un fichier séparé les métadonnées des éléments enregistrés sur le support de données. Nous avons décidé d'y accoler systématiquement le terme archivistique auquel cette fonctionnalité informatique se rattache dans le monde analogique, le « récolement », compris de manière générale comme l'« opération consistant à dresser la liste topographique des articles conservés dans un service d'archives ou un fonds. Désigne aussi l'opération destinée à vérifier l'intégralité des fonds et collections d'un service d'archives périodiquement ou lors du changement de responsable d'un service d'archives. » (Portail International Archivistique Francophone (PIAF) 2015b).

<sup>5</sup> Le terme arborescence dont il a déjà été question est compris de la manière suivante dans la suite de ce travail : une « représentation organisationnelle adoptant la forme d'un arbre, qui établit une stricte hiérarchie entre les éléments qui la composent, de façon que toute

redondances strictes » ; et « Comparaison de données ». Nous verrons au fil de cette étude que ces fonctionnalités peuvent être réparties en deux groupes : les fonctionnalités que l'on pourrait appeler « descriptives » (« Extraction de métadonnées / Récolement » et « Analyse de l'arborescence et de la volumétrie »), c'est-à-dire celles qui vont nous permettre de prendre connaissance du contenu des supports, de comprendre comment ce dernier est organisé (de sorte par exemple à pouvoir ensuite mener une évaluation *top-down* par séries et dossiers) et de documenter l'acquisition et le résultat des actions entreprises ; et les fonctionnalités dites « actives » (« Recherche de redondances strictes » et « Comparaison de données »), puisque ces dernières engendrent la manipulation de données et en particulier leur élimination.

### 3.1.2 Critères de sélection des outils testés

Au vu du très grand nombre d'outils référencés par les différents organismes et rapports archivistiques, il est nécessaire de sélectionner les logiciels qui feront l'objet d'un test en situation réelle avec les données de la Cinémathèque. Le choix des logiciels à tester s'est donc appuyé sur les critères suivants :

**Fonctionnalités.** Le logiciel doit proposer au moins une des fonctionnalités présentées précédemment, que ce soit une fonctionnalité *descriptive*, permettant la prise de connaissance du contenu des supports, ou une fonctionnalité *active*, permettant la manipulation et le traitement des données.

**Prix.** Le prix d'achat ou d'utilisation du logiciel doit être relativement faible, et ne doit pas être un frein à l'acquisition d'un logiciel de qualité par la Cinémathèque. On veillera à conseiller des logiciels gratuits, ainsi que des outils n'exigeant pas des licences utilisateur multiples et coûteuses ou des frais d'abonnement excessifs ; les logiciels payants proposant des périodes d'essai avant acquisition seront privilégiés.

**Accessibilité technique.** Les logiciels libres (*open source software*, distribués par exemple avec des licences *GNU General Public License*, *BSD Berkeley Software Distribution License* ou *The MIT License*) seront avantagés, c'est-à-dire ceux dont le code source est partagé, et qui peut être modifié et réutilisé. Cette accessibilité technique permet non seulement de garantir la transparence sur le fonctionnement des outils (diminuant l'aspect *boîte noire* qui entoure encore certains algorithmes par exemple), mais également d'adapter l'outil aux besoins spécifiques de l'institution et de participer à son développement au sein d'une communauté d'utilisateurs.

**Accessibilité cognitive.** Le logiciel doit être accessible au plus grand nombre, et tout particulièrement aux archivistes ne possédant pas des connaissances informatiques avancées. Il devra donc présenter une interface graphique (*GUI*, *Graphical User Interface*) – ce qui exclut les outils en ligne de commande (*CLI*, *Command Line Interface*) –, et une ergonomie permettant une prise en main aisée. En outre, on veillera à sélectionner des outils offrant des manuels d'utilisation complet, voire des forums de discussion pour favoriser les échanges entre utilisateurs et les retours d'expérience.

### 3.1.3 Liste des outils testés

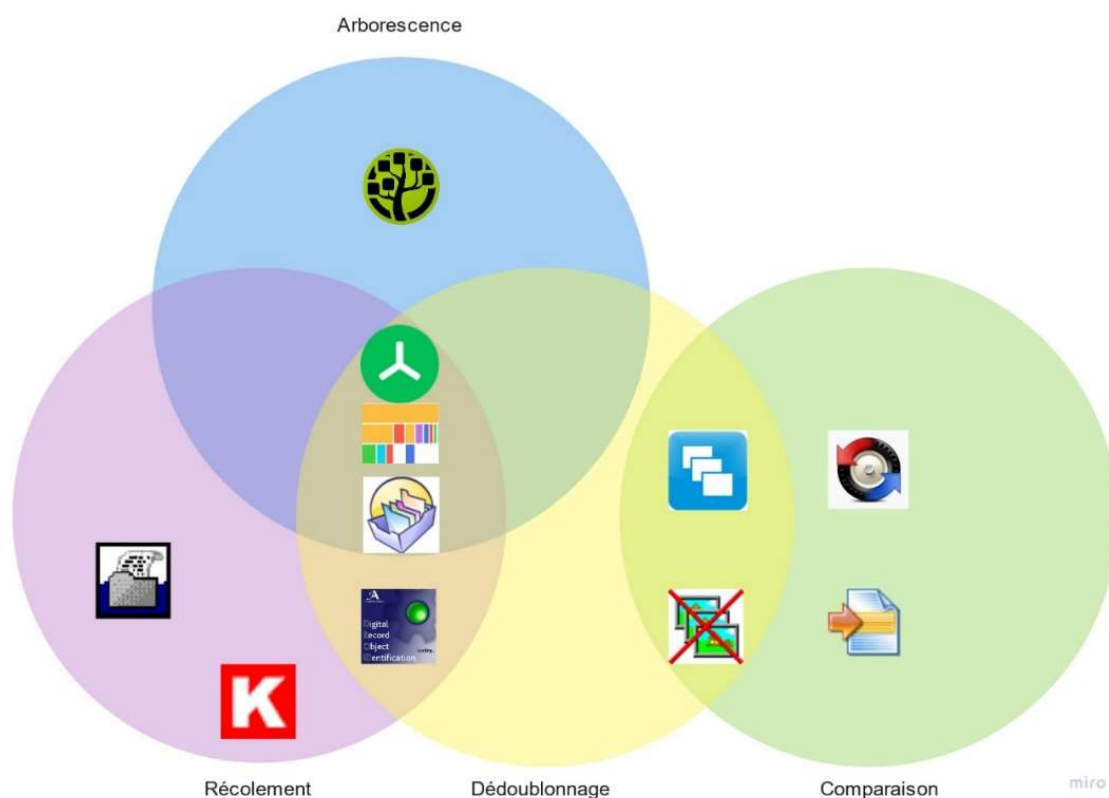
De nombreux logiciels répondent aux critères présentés ci-dessus. Pour réduire la liste des candidats et affiner la sélection, nous avons opéré un choix par « famille de logiciels par

---

information, sauf la première, procède d'une seule autre, mais peut en engendrer plusieurs. »  
(Grand dictionnaire terminologique 2012)

fonctionnalité », et avons sélectionné *au moins* deux outils par famille. Cette méthode permet non seulement les comparaisons *inter*-logiciels mais aussi d'avoir une représentation plus complète des fonctionnalités offertes et de la manière dont une fonctionnalité est exécutée et mise en pratique. Le test d'un seul outil par famille pourrait effectivement introduire un biais important en associant trop étroitement une fonctionnalité et son exécution particulière par un logiciel. Les familles de logiciels recoupent peu ou prou les fonctionnalités analysées : « Arborescence et volumétrie », « Extraction de métadonnées / Récolement », « Recherche de redondances strictes », « Comparaison de données ». Ainsi, à titre d'exemple, pour ce qui concerne la fonctionnalité « Extraction de métadonnées / Récolement », deux logiciels n'offrant que cette fonctionnalité ont été testés (*Karen's Directory Printer* et *Pir*) aux côtés d'autres logiciels plus généralistes offrant la possibilité d'extraire des métadonnées parmi leurs fonctionnalités.

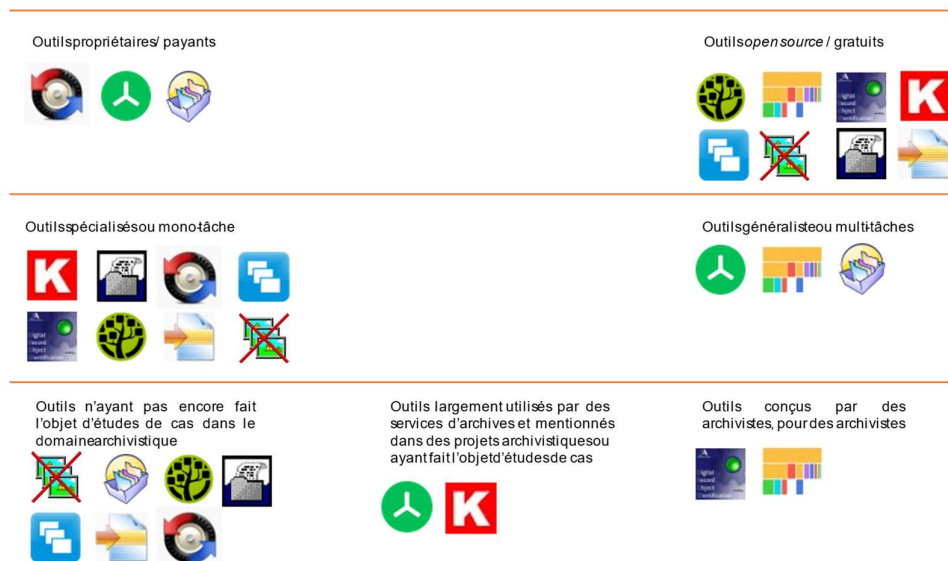
Figure 6 : Familles de logiciels par fonctionnalité



Le choix définitif des logiciels testés a également été déterminé de manière à avoir un panel aussi diversifié et représentatif que possible. Pour chaque famille de fonctionnalités, nous avons ainsi essayé, dans la mesure du possible, de choisir des logiciels aux caractéristiques différentes et provenant d'environnements de développement distincts : les logiciels *open source* côtoient des logiciels propriétaires et payants proposant des essais gratuits ; les logiciels généralistes ou multi-tâches figurent aux côtés de logiciels spécialisés ou mono-tâche ; et des outils pensés *par* des archivistes *pour* des archivistes sont testés en parallèle avec des outils qui ont été développés dans des domaines d'expertise éloignés du domaine des *GLAM* (*Galleries, Libraries, Archives and Museums*), et qui sont utilisés plutôt par des services informatiques et dans le domaine de la bureautique.




Figure 7 : Types de logiciels (caractéristiques et développements)




Finalement, le nombre total d'outils testés a été déterminé en fonction des limites imposées par la nature académique de cette recherche (temps limité à disposition pour la rédaction d'un Mémoire de Master) et l'échantillon a été délimité de sorte à atteindre une forme de saturation – quelques outils, non listés ci-dessous, ont été testés superficiellement afin de garantir que des fonctionnalités essentielles n'aient pas été négligées (c'est par exemple le cas, dans le domaine du dédoublement, des logiciels *CloneSpy* et *Duplicate Finder* largement cités dans la littérature ou, pour ce qui concerne l'étude de l'arborescence et de la volumétrie, par les logiciels *WizTree* – très proche dans ses fonctionnalités et son interface de *WinDirStat* – et *FolderSizes* – qui peut être assimilé à l'outil *TreeSize Professional* sur bien des aspects – qui nous ont été conseillés par le service informatique de la Cinémathèque suisse).

Fort de tous ces critères, le choix des logiciels testés est donc le suivant : *AllDup*, *AntiDupl*, *Archifiltre*, *Beyond Compare*, *Droid*, *Karen's Directory Printer*, *Pir*, *TreeSize Professional*, *WinCatalog*, *WinDirStat* et *WinMerge*. Pour chacun des logiciels cités, nous avons établi une fiche récapitulative (voir 3.1.3.1 à 3.1.3.11) qui présente sommairement les caractéristiques de l'outil selon les informations données par les développeurs sur leurs sites internet ou sur les pages *GitHub* dédiées aux logiciels *open source* – à ce stade, nous avons majoritairement repris les formulations proposées par les développeurs, en particulier pour ce qui est des fonctionnalités générales, en menant parfois un important travail de sélection pour restreindre des listes extrêmement longues, (trop) détaillées et répétitives (rappelons que certains de ces outils sont payants et que la liste des fonctionnalités fait alors partie des arguments de vente).


### 3.1.3.1 AllDup

 <b>Présentation</b>	<i>Nom</i>	AllDup
	<i>Nom complet</i>	-
	<i>Développeur</i>	MTSD (Chonburi, Thaïlande)
	<i>Présentation par le développeur</i>	« AllDup is a freeware tool for searching and removing file duplicates on your computer, network shares or external storage media. The fast search algorithm find duplicates of any file type, e.g., text, pictures, music or movies. The powerful search engine enables you to find duplicates with a combination of the following criteria: File Name, File Extension, File Size, File Content, File Dates and File Attributes. Additionally you can also search for similar file names, similar or almost identical pictures and similar or almost identical music files. Furthermore, you can find video & audio files with the same or almost same audio length or search your hard disk for Hard Links. »
	<i>Public cible visé</i>	« For more than 20 years, MTSD's focus has been developing innovative software products for business and home users. »
	<i>Slogan</i>	« Find and remove duplicate files »
	<i>Site internet (URL principale)</i>	<a href="https://www.allsync.biz/alldup_help/alldup.php">https://www.allsync.biz/alldup_help/alldup.php</a> (consulté le 20 juin 2022)
	<i>Version</i>	4.5.16
<b>Fonctionnalités</b>	<i>Fonctionnalités proposées par le développeur (sélection)</i>	Find duplicate files with a combination of the following criteria : file content, file name, file extension, file dates or file attributes
		Find, remove, delete, copy or move duplicate files
		Find similar pictures
		Find similar music files
		Find files with similar names
		Create shortcuts or hard links to the last original file
<b>Dates</b>	<i>Date de la première version</i>	[2000]
	<i>Date de la dernière mise à jour</i>	08.03.2022
<b>Accessibilité</b>	<i>Distribution (licence)</i>	Freeware
	<i>Installation</i>	Client (il existe une version portable)
	<i>Prix</i>	-
	<i>Essai gratuit (durée)</i>	-
<b>Caractéristiques techniques</b>	<i>Configuration informatique requise</i>	AllDup is compatible with all versions of the following operating systems (32/64-bit): Workstation : Microsoft Windows 7, 8, 10, 11, Vista*, XP* Server: Microsoft Windows Server 2003*, 2008*, 2012, 2016, 2019, 2022 (* Please use the AllDup Portable Edition if you want to use AllDup with these older Windows operating systems. The AllDup setup file can only be installed from Windows 7 / Server 2012.)
	<i>Poids</i>	20.50 MB
	<i>Écrit en</i>	-
<b>Utilisation</b>	<i>Manuel d'utilisation</i>	Manuel d'utilisation téléchargeable en PDF et CHM : <a href="https://www.alldup.de/en_download_alldup.php">https://www.alldup.de/en_download_alldup.php</a> ( <a href="https://www.alldup.de/en_download_alldup.php#pdf">https://www.alldup.de/en_download_alldup.php#pdf</a> <a href="https://www.alldup.de/en_download_alldup.php#chm">https://www.alldup.de/en_download_alldup.php#chm</a> )
	<i>Langue(s)</i>	English, German, Dutch, Russian, Ukrainian, Chinese, Spanish, Thai, French, Hungarian, Polish, Italian, Swedish, Arabic, Portuguese, Slovenian, Finnish, Greek, Czech, Turkish
	<i>Étude(s) de cas publiée(s)</i>	-
	<i>Source(s)</i>	Benedetti 2021 ; COPTR ; Nestor

### 3.1.3.2 AntiDupl

 <b>Présentation</b>	<i>Nom</i>	AntiDupl
	<i>Nom complet</i>	AntiDupl.NET
	<i>Développeur</i>	Yermalayeu Ihar, Borisov Dmitry
	<i>Présentation par le développeur</i>	« Typically, modern computer users have large collections of images in various formats. And then more these collections, then more likely to have the large number of duplicates. The natural desire of the user is to get rid of them. However, if the collection is large enough, do this manually is a very tedious and unproductive work. AntiDupl.NET program will help you automate this process. It can find and display duplicate images in the main graphic formats: JPEG, GIF, TIFF, BMP, PNG, EMF, WMF, EXIF, ICON, JP2, PSD, DDS and TGA. The comparison is based on the contents of the files, so the program can find not only almost identical, but similar images. In addition, the program can find images with some types of defects. AntiDupl.NET program is free and open-source software. It is simple to use, has high speed and accuracy of work, supports Russian, Byelorussian, German and English interface. »
	<i>Public cible visé</i>	-
	<i>Slogan</i>	« Search of similar and defective images on the disk »
	<i>Site internet (URL principale)</i>	<a href="https://ermig1979.github.io/AntiDupl/english/index.html">https://ermig1979.github.io/AntiDupl/english/index.html</a> (consulté le 23.07.2022)
	<i>Version</i>	AntiDupl.NET-2.3.10
	<i>Fonctionnalités proposées par le développeur</i>	Can find and display duplicate images Can find not only almost identical, but similar images Can find images with some types of defects
<b>Dates</b>	<i>Date de la première version</i>	17.10.2003
	<i>Date de la dernière mise à jour</i>	03.03.2020
<b>Accessibilité</b>	<i>Distribution (licence)</i>	Licence MIIT Free and open-source software Copyright © 2002-2018, Yermalayeu Ihar
	<i>Installation</i>	Client
	<i>Prix</i>	-
	<i>Essai gratuit (durée)</i>	-
<b>Caractéristiques techniques</b>	<i>Configuration informatique requise</i>	Windows Vista, Windows 7, Windows 8 ; Windows 2000 Service Pack 3, Windows Server 2003, and Windows XP Service Pack 2 with the installation of Microsoft.NET Framework 2.0 (22 Mb) or higher ; For Windows 2000 may need to add to the system directory file gdiplus.dll.
	<i>Poids</i>	3.6 MB
	<i>Écrit en</i>	C++
<b>Utilisation</b>	<i>Manuel d'utilisation</i>	<a href="https://ermig1979.github.io/AntiDupl/data/help/english/index.html">https://ermig1979.github.io/AntiDupl/data/help/english/index.html</a> (consulté le 29.05.2022)
	<i>Langue(s)</i>	Russian [Byelorussian, German] and English interface
	<i>Étude(s) de cas publiée(s)</i>	-
	<i>Source(s)</i>	Christophe Uldry (Cinémathèque suisse, Secteur Iconographie)

### 3.1.3.3 Archifiltre


 <b>Présentation</b>	<i>Nom</i>	Archifiltre
	<i>Nom complet</i>	Docs par Archifiltre
	<i>Développeur</i>	Fabrique Numérique des Ministères Sociaux (start-up d'État française faisant partie de la)
	<i>Présentation par le développeur</i>	« Archifiltre est un logiciel libre d'analyse et de traitement d'arborescences de fichiers bureautiques non-structurés, développé par les ministères sociaux. Son objectif est de proposer, à tout utilisateur de fichiers bureautiques, un outil de visualisation d'arborescences complètes afin de pouvoir les analyser, les auditer, les trier, les enrichir et les verser dans un système d'archivage électronique (SAE). »
	<i>Public cible visé</i>	« Docs peut être utilisé par tout le monde. Certaines fonctionnalités ont été conçues spécialement pour les archivistes, mais le logiciel peut être utilisé par tout le monde pour visualiser des arborescences de fichiers, auditer, et enrichir leurs métadonnées. »
	<i>Slogan</i>	« Visualisez et améliorez vos arborescences de fichiers ! »
	<i>Site internet (URL principale)</i>	<a href="https://archifiltre.fabrique.social.gouv.fr/">https://archifiltre.fabrique.social.gouv.fr/</a> (consulté le 11 juillet 2022)
	<i>Version</i>	v3.2.2 Romantic Raccoon
<b>Fonctionnalités</b>		Appréhender des arborescences
	<i>Fonctionnalités proposées par le développeur</i>	Enrichir des métadonnées Mener une opération d'audit Identifier les redondances
	<i>Date de la première version</i>	02.03.2018
	<i>Date de la dernière mise à jour</i>	22.09.2021
<b>Accessibilité</b>	<i>Distribution (licence)</i>	Libre et gratuit Licence MIT ( <a href="https://fr.wikipedia.org/wiki/Archifiltre">https://fr.wikipedia.org/wiki/Archifiltre</a> ) consulté le 11 juillet 2022)
	<i>Installation</i>	Application locale (installation non nécessaire) Application web jusqu'à la version 1.8 (versions plus maintenues à jour)
	<i>Prix</i>	-
	<i>Essai gratuit (durée)</i>	-
<b>Caractéristiques techniques</b>	<i>Configuration informatique requise</i>	Système d'exploitation Windows 64 bits, Windows 32 bits, Linux ou MacOS ; 8 Go de RAM ; pas de limite de mémoire de processus
	<i>Poids</i>	63, 4 Mo
	<i>Écrit en</i>	Javascript et Typescript
<b>Utilisation</b>	<i>Manuel d'utilisation</i>	<a href="https://github.com/SocialGouv/archifiltre-docs/wiki/Wiki-Archifiltre">https://github.com/SocialGouv/archifiltre-docs/wiki/Wiki-Archifiltre</a> (consulté le 11 juillet 2022)
	<i>Langue(s)</i>	Français ; Anglais ; Allemand
	<i>Étude(s) de cas publiée(s)</i>	Fritz 2021
	<i>Source(s)</i>	Benedetti 2021 ; COPTR ; Naud 2019 ; Nestor

### 3.1.3.4 Beyond Compare


 <b>Présentation</b>	<i>Nom</i>	Beyond Compare
	<i>Nom complet</i>	-
	<i>Développeur</i>	Scooter Software
	<i>Présentation par le développeur</i>	« Focused. Intelligent Comparison. Compare files and folders using simple, powerful commands that focus on the differences you're interested in and ignore those you're not. Merge changes, synchronize files, and generate reports. Agile. Access Data Anywhere. Directly access FTP sites, media devices, WebDAV resources, svn repositories and cloud storage. All from your Windows, macOS or Linux workstation. Multifaceted. Specialized Viewers. Beyond Compare includes built-in comparison viewers for a variety of data types. In addition to text, compare tables, images, binary files, registry hives, and much more. »
	<i>Public cible visé</i>	-
	<i>Slogan</i>	« Reconcile your Differences »
	<i>Site internet (URL principale)</i>	<a href="https://www.scootersoftware.com/">https://www.scootersoftware.com/</a> (consulté le 09.06.2022)
	<i>Version</i>	4.4.2.
<b>Fonctionnalités</b>	<i>Fonctionnalités proposées par le développeur (sélection)</i>	Compare files
		Text Compare
		Table Compare
		Binary Compare
		Registry Compare
		Picture Compare
<b>Dates</b>	<i>Date de la première version</i>	26.06.1996
	<i>Date de la dernière mise à jour</i>	16.03.2022
<b>Accessibilité</b>	<i>Distribution (licence)</i>	Propriétaire
	<i>Installation</i>	Client
	<i>Prix</i>	Pro Edition : 60 \$ ; Standard Edition : 30 \$. « Beyond Compare is licensed on either a per-user or per-workstation basis. A single user license covers one person using BC on any number of computers, or a single workstation accessed by multiple people. [...] Beyond Compare licenses are perpetual. Once you buy a license to a specific major version, and as long as you abide by the license agreement, you can use that version forever with no additional cost. »
	<i>Essai gratuit (durée)</i>	30 jours d'utilisation effective
<b>Caractéristiques techniques</b>	<i>Configuration informatique requise</i>	Windows, macOS, Linux ; Intel processor; 1 GHz or faster recommended ; 1 GB RAM; additional memory recommended for large comparisons ; 50 MB of available hard drive space ; 1024 x 768 display resolution
	<i>Poids</i>	22402 kb
	<i>Écrit en</i>	-
<b>Utilisation</b>	<i>Manuel d'utilisation</i>	<a href="https://www.scootersoftware.com/v4help/">https://www.scootersoftware.com/v4help/</a> (consulté le 23.07.2022)
	<i>Langue(s)</i>	Anglais, Allemand, Français, Japonais, Chinois
	<i>Étude(s) de cas publiée(s)</i>	Shein 2014
	<i>Source(s)</i>	KOST/CECO




### 3.1.3.5 Droid

 <b>Présentation</b>	<i>Nom</i>	DROID
	<i>Nom complet</i>	Digital Record Object Identification
	<i>Développeur</i>	The National Archives UK
	<i>Présentation par le développeur</i>	« DROID is a software tool developed by The National Archives to perform automated batch identification of file formats. Developed by our Digital Preservation department as part of its broader digital preservation activities, DROID is designed to meet the fundamental requirement of any digital repository to be able to identify the precise format of all stored digital objects, and to link that identification to a central registry of technical information about that format and its dependencies. » (site internet)
	<i>Public cible visé</i>	« It is in widespread use across the world, in cultural memory institutions, local and central government departments and other public bodies, and has been embedded into multiple commercial and open source digital preservation products. » (DROID : User Guide 2020, 3)
	<i>Slogan</i>	« file format identification tool »
	<i>Site internet (URL principale)</i>	<a href="https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-">https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-</a> (consulté le 12 mai 2022)
	<i>Version</i>	DROID 6.5.2
<b>Fonctionnalités</b>	<i>Fonctionnalités proposées par le développeur (sélection)</i>	The core function of DROID is accurate file format identification, even if the file extension is wrong or missing. In addition to identifying the file format DROID also extracts other information about the files it scans, such as, file size, last modified date and file path.
<b>Dates</b>	<i>Date de la première version</i>	[2005]
	<i>Date de la dernière mise à jour</i>	01.05.2020
<b>Accessibilité</b>	<i>Distribution (licence)</i>	Free and open source software New BSD License
	<i>Installation</i>	Client
	<i>Prix</i>	-
	<i>Essai gratuit (durée)</i>	-
<b>Caractéristiques techniques</b>	<i>Configuration informatique requise</i>	« The DROID version without embedded Java runs on any platform with Java 8 to 11 Standard Edition (SE) or OpenJDK installed. A Windows platform is required for the option with embedded Java. DROID is built and tested on: Linux CentOS (Red Hat), Microsoft Windows 10 (64 bit), Raspbian, Mac OSX (Mojave), Mac OSX (Sierra), Linux Ubuntu Desktop »
	<i>Poids</i>	88.2 MB (zip)
	<i>Écrit en</i>	Java
<b>Utilisation</b>	<i>Manuel d'utilisation</i>	En ligne et en pdf <a href="https://www.nationalarchives.gov.uk/documents/information-management/droid-user-guide.pdf">https://www.nationalarchives.gov.uk/documents/information-management/droid-user-guide.pdf</a> Google Groups discussion ( <a href="https://groups.google.com/g/droid-list">https://groups.google.com/g/droid-list</a> ) et courriel (PRONOM@nationalarchives.gov.uk) en cas de questions ou de besoin d'assistance ; rapport d'erreur, contribution au code source ou fonctionnalité souhaitée via la page GitHub Issues.
	<i>Langue(s)</i>	Anglais
	<i>Étude(s) de cas publiée(s)</i>	Sloyan 2016
	<i>Source(s)</i>	Benedetti 2021 ; COPTR ; KOST-CECO ; Nestor

### 3.1.3.6 Karen's Directory Printer

 <b>Présentation</b>	<i>Nom</i>	Karen's Directory Printer
	<i>Nom complet</i>	Karen's Directory Printer
	<i>Développeur</i>	Karen's Power Tools (@Karen Kenworthy, Joe Winett)
	<i>Présentation par le développeur</i>	« No more fumbling with My Computer or Windows Explorer, wishing you could print information about all your files. Karen's Directory Printer can print the name of every file on a drive, along with the file's size, date and time of last modification, and attributes (Read-Only, Hidden, System and Archive)! And now, the list of files can be sorted by name, size, date created, date last modified, or date of last access. » (site internet)
	<i>Public cible visé</i>	-
	<i>Slogan</i>	« the File Cataloging Utility for Windows »
	<i>Site internet (URL principale)</i>	<a href="https://www.karenware.com/powertools/karens-directory-printer">https://www.karenware.com/powertools/karens-directory-printer</a> (consulté le 11 mai 2022)
	<i>Version</i>	5.4.4.
<b>Fonctionnalités</b>	<i>Fonctionnalités proposées par le développeur</i>	This Power Tool can print the names of every file in a folder or drive. It can also print a file or sub-folder's size, date of creation, date of last modification, date of last access, attributes (read-only, system, hidden, compressed, etc.) and more! Additional options allow the information to be saved to a disk file instead of printed. Once stored on disk, the folder and file information can be e-mailed, imported into a spreadsheet or other program, or saved for future analysis and comparison.
<b>Dates</b>	<i>Date de la première version</i>	[1997]
	<i>Date de la dernière mise à jour</i>	20.05.2020
<b>Accessibilité</b>	<i>Distribution (licence)</i>	These programs are licensed for free for personal and educational use. However, commercial use (any use at a business or on your job) requires an inexpensive license.
	<i>Installation</i>	Client
	<i>Prix</i>	15 \$ pour une licence individuelle ; tarif dégressif pour les licences d'entreprise, par nombre total de licences.
	<i>Essai gratuit (durée)</i>	-
<b>Caractéristiques techniques</b>	<i>Configuration informatique requise</i>	Karen's Power Tools are compatible with virtually all versions of Windows released since Windows 95.
	<i>Poids</i>	1773936 bytes
	<i>Écrit en</i>	-
<b>Utilisation</b>	<i>Manuel d'utilisation</i>	En ligne, menu d'aide du logiciel
	<i>Langue(s)</i>	Anglais
	<i>Étude(s) de cas publiée(s)</i>	Paradigm 2007 ; AIMS 2012 ; Shein 2014
	<i>Source(s)</i>	COPTR

### 3.1.3.7 Pir


 <b>Présentation</b>	<i>Nom</i>	Pir
	<i>Nom complet</i>	-
	<i>Développeur</i>	PEK's Productions (Suède)
	<i>Présentation par le développeur</i>	« Use Pir to create lists of files and folders. There are several settings to use to create lists that have the information that you need. You can, for example, list size of files and folders, checksums (CRC32 and MD5), file types and much more. »
	<i>Public cible visé</i>	-
	<i>Slogan</i>	« Simple application to create lists of files and folders »
	<i>Site internet (URL principale)</i>	<a href="https://www.pekspro.com/products/pir/">https://www.pekspro.com/products/pir/</a> (consulté le 13.08.2022)
	<i>Version</i>	1.11
<b>Fonctionnalités</b>	<i>Fonctionnalités proposées par le développeur</i>	Create lists of files and folders
<b>Dates</b>	<i>Date de la première version</i>	[1997]
	<i>Date de la dernière mise à jour</i>	11.07.2004
<b>Accessibilité</b>	<i>Distribution (licence)</i>	Gratuit
	<i>Installation</i>	Client
	<i>Prix</i>	-
	<i>Essai gratuit (durée)</i>	-
<b>Caractéristiques techniques</b>	<i>Configuration informatique requise</i>	Windows
	<i>Poids</i>	148 Ko
	<i>Écrit en</i>	-
<b>Utilisation</b>	<i>Manuel d'utilisation</i>	-
	<i>Langue(s)</i>	Anglais
	<i>Étude(s) de cas publiée(s)</i>	-
	<i>Source(s)</i>	Benedetti 2021




### 3.1.3.8 TreeSize Professional

 <b>Présentation</b>	<i>Nom</i>	TSP
	<i>Nom complet</i>	TreeSize Professional
	<i>Développeur</i>	Jam Software (Trier, Allemagne)
	<i>Présentation par le développeur</i>	« As multifunctional as a Swiss Army knife is the market-leading file and disk space manager TreeSize. The software analyses all stored data across your systems and visualizes the results in meaningful charts and statistics. Find out where your disk space has gone at a glance and take immediate action if necessary. For this purpose, TreeSize provides you with a wide range of file management options. With our all-round performer you have a multi-tool in your hand to organize your storage systems and to get your valuable storage space back. The world-renowned solution is optimized for handling large file volumes and complements Windows Explorer. A high degree of automation is enabled by command line parameters and management of scans scheduling directly in a comfortable, graphical user interface. » (site internet)
	<i>Public cible visé</i>	-
	<i>Slogan</i>	« The Powerful Graphical Manager For Your Storage Systems »
	<i>Site internet (URL principale)</i>	<a href="https://www.jam-software.com/treesize">https://www.jam-software.com/treesize</a> (consulté le 25.07.2022)
	<i>Version</i>	8.3.2.1665
	<i>Fonctionnalités proposées par le développeur (sélection)</i>	Manage and clean up disk space efficiently Visualize disk usage Analyze in detail, down to all directory levels Find and remove redundant files Numerous export and reporting possibilities
<b>Dates</b>	<i>Date de la première version</i>	1997
	<i>Date de la dernière mise à jour</i>	11.2021 (v8)
<b>Accessibilité</b>	<i>Distribution (licence)</i>	Propriétaire
	<i>Installation</i>	Client
	<i>Prix</i>	47,95 € pour une licence <i>TreeSize Professional</i> (inclus 12 mois de support et de mise à jour) ; tarif dégressif par nombre total de licences. « All our licenses are perpetual. »
	<i>Essai gratuit (durée)</i>	30 jours dès l'installation
<b>Caractéristiques techniques</b>	<i>Configuration informatique requise</i>	TreeSize requires Windows 8.1, Server 2012, or upwards as well as .Net framework 4.8
	<i>Poids</i>	27.82 MB
	<i>Écrit en</i>	-
<b>Utilisation</b>	<i>Manuel d'utilisation</i>	En ligne et en pdf <a href="https://www.jam-software.com/treesize/help.shtml">https://www.jam-software.com/treesize/help.shtml</a> (consulté le 25.07.2022)
	<i>Langue(s)</i>	English, German, Chinese (simplified), Czech, Danish, Dutch, French, Greek, Japanese, Korean, Portuguese, Russian, Slovenian, Spanish, Ukrainian.
	<i>Étude(s) de cas publiée(s)</i>	Belovari 2017 ; Birn 2017 ; Fritz 2021
	<i>Source(s)</i>	COPTR ; KOST-CECO ; Nestor


### 3.1.3.9 WinCatalog

 <b>Présentation</b>	<i>Nom</i>	WinCatalog 2021
	<i>Nom complet</i>	WinCatalog 2021 - Disk Catalog Software for Windows
	<i>Développeur</i>	OrangeCat Software, LLC
	<i>Présentation par le développeur</i>	« Automatically create a catalog of all files, stored on your disks (HDDs, DVDs, CDs, network drives and other media storage devices): WinCatalog will automatically grab ID3 tags for music files, Exif tags and thumbnails for photos, thumbnails and basic information for video files, e-books, contents of archive files, thumbnails for images (pictures) and PDF files, ISO files, and much more. Organize your file catalog, using virtual folders, tags (categories) and user defined fields, and find files in seconds, using powerful search, even when disks are not connected to the computer. Also easily use WinCatalog as a duplicate file finder. Your disk catalog can be automatically updated through Windows task scheduler. »
	<i>Public cible visé</i>	-
	<i>Slogan</i>	« Catalogs and organizes disks, files, and folders »
	<i>Site internet (URL principale)</i>	<a href="https://www.wincatalog.com/">https://www.wincatalog.com/</a> (consulté le 06.08.2022)
	<i>Version</i>	.2021.2.1
<b>Fonctionnalités</b>	<i>Fonctionnalités proposées par le développeur (sélection)</i>	An ability to add disks, folders, and individual files to your catalog
		All types of files are indexed, including extracting additional information from various types of files (archives, photos, images, video, music, e-books, PDF, html, txt)
		Adding virtual folders and moving items between them
		Managing tags and associating them with individual files, disks, or folders
		Duplicate file finder - an ability to find duplicate files by matching one or more fields or a check sum
		Exporting your catalog into XML, HTML or CSV report (e.g. for export to MS Excel or other applications) and printing
<b>Dates</b>	<i>Date de la première version</i>	2001
	<i>Date de la dernière mise à jour</i>	08.04.2022
<b>Accessibilité</b>	<i>Distribution (licence)</i>	Propriétaire
	<i>Installation</i>	Client
	<i>Prix</i>	29.95 € WinCatalog 2021 Personal 49.95 € WinCatalog 2021 Professional (« Both types of licenses include the same features and allow installation on any number of computers per a licens », « Unlimited installations and activations »)
	<i>Essai gratuit (durée)</i>	30 jours
<b>Caractéristiques techniques</b>	<i>Configuration informatique requise</i>	Windows XP SP3 Windows Vista (all versions) Windows 7 (all versions) Windows 8 (runs as desktop application) Windows 8.1 (runs as desktop application) Windows 10 (runs as desktop application) Windows 11 (runs as desktop application) Microsoft Visual C++ 2008 Redistributable Package .NET Framework 4.8
	<i>Poids</i>	[87 Mo]
	<i>Écrit en</i>	C++
<b>Utilisation</b>	<i>Manuel d'utilisation</i>	Oui, en ligne et en version PDF <a href="https://www.wincatalog.com/feedback.html">https://www.wincatalog.com/feedback.html</a> (consulté le 25.05.2022)
	<i>Langue(s)</i>	English (default), Český, Deutsch, Español Europeo & Español Latino, Français, Italiano, Magyar, Nederlands, Polski, Português Brasileiro & Portugal, Română, Slovak, Slovenski, Suomi, Svenska, Türkçe (etc.)
	<i>Étude(s) de cas publiée(s)</i>	Kim 2006
	<i>Source(s)</i>	-

### 3.1.3.10 WinDirStat

 <b>Présentation</b>	<i>Nom</i>	WinDirStat
	<i>Nom complet</i>	Windows Directory Statistics
	<i>Développeur</i>	Bernhard Seifert, Oliver Schneider
	<i>Présentation par le développeur</i>	« WinDirStat is a disk usage statistics viewer and cleanup tool for various versions of Microsoft Windows. »
	<i>Public cible visé</i>	-
	<i>Slogan</i>	« Shows where all your disk space has gone, and helps you clean it up. »
	<i>Site internet (URL principale)</i>	<a href="https://windirstat.net/">https://windirstat.net/</a> (consulté le 25.07.2022)
	<i>Version</i>	WinDirStat 1.1.2.80 (Unicode)
<b>Fonctionnalités</b>	<i>Fonctionnalités proposées par le développeur (sélection)</i>	The directory list, which resembles the tree view of the Windows Explorer but is sorted by file/subtree size
		The treemap, which shows the whole contents of the directory tree straight away
		The extension list, which serves as a legend and shows statistics about the file types.
<b>Dates</b>	<i>Date de la première version</i>	[2003]
	<i>Date de la dernière mise à jour</i>	-
<b>Accessibilité</b>	<i>Distribution (licence)</i>	Open Source Software, GNU Public License, version 2 (GPLv2)
	<i>Installation</i>	Client
	<i>Prix</i>	-
	<i>Essai gratuit (durée)</i>	-
<b>Caractéristiques techniques</b>	<i>Configuration informatique requise</i>	Windows 95 (IE5), Windows 98 SE, Windows ME, Windows NT4 (SP5), Windows 2000, Windows XP, Windows Vista, Windows 7, 8 and 8.1.
	<i>Poids</i>	636 Ko
	<i>Écrit en</i>	C++
<b>Utilisation</b>	<i>Manuel d'utilisation</i>	Menu d'aide intégré au logiciel ; une version PDF en également disponible : <a href="https://documentation.help/WinDirStat/documentation.pdf">https://documentation.help/WinDirStat/documentation.pdf</a>
	<i>Langue(s)</i>	Czech, Dutch, Estonian, Finnish, German, English, Spanish, French, Hungarian, Italian, Polish, Russian
	<i>Étude(s) de cas publiée(s)</i>	-
	<i>Source(s)</i>	Benedetti 2021

### 3.1.3.11 WinMerge

 <b>Présentation</b>	<i>Nom</i>	WinMerge
	<i>Nom complet</i>	WinMerge
	<i>Développeur</i>	-
	<i>Présentation par le développeur</i>	« WinMerge est un outil Open Source de différenciation et de fusion pour Windows. WinMerge peut comparer et les dossiers et les fichiers, en représentant les différences dans un fichier texte visuel qui est facile à comprendre ou manipuler. [...] WinMerge est très utile pour déterminer ce qui a changé entre les versions d'un projet, et fusionner les changements entre les versions. WinMerge peut être utilisé comme outil externe de comparaison / fusion ou en tant qu'application autonome. De plus, Winmerge a de nombreuses fonctionnalités utiles qui rendent les comparaisons, la synchronisation, et la fusion aussi faciles et efficaces que possible »
	<i>Public cible visé</i>	-
	<i>Slogan</i>	« You will see the difference... »
	<i>Site internet (URL principale)</i>	<a href="https://winmerge.org/">https://winmerge.org/</a> (consulté le 26 juillet 2022)
	<i>Version</i>	WinMerge 2.16.20
<b>Fonctionnalités</b>	<i>Fonctionnalités proposées par le développeur (sélection)</i>	Comparaison de fichiers
		Comparaison de dossiers
		Comparaison d'images
		Comparaison de tableurs
		Contrôle de versions
<b>Dates</b>	<i>Date de la première version</i>	[2000]
	<i>Date de la dernière mise à jour</i>	27.04.2022
<b>Accessibilité</b>	<i>Distribution (licence)</i>	Open Source sous Licence Publique Générale GNU.
	<i>Installation</i>	Client
	<i>Prix</i>	-
	<i>Essai gratuit (durée)</i>	-
<b>Caractéristiques techniques</b>	<i>Configuration informatique requise</i>	32-bit installer: Microsoft Windows XP SP3 or newer 64-bit installer: Microsoft Windows 7 or newer
	<i>Poids</i>	8,17 MB
	<i>Écrit en</i>	C++
	<i>Manuel d'utilisation</i>	Disponible en ligne, sur le site internet (en anglais et en japonais) : <a href="https://manual.winmerge.org/en/">https://manual.winmerge.org/en/</a>
<b>Utilisation</b>	<i>Langue(s)</i>	Arabic, Basque, Brazilian, Bulgarian, Catalan, Chinese (Simplified), Chinese (Traditional), Corsican, Croatian, Czech, Danish, Dutch, English, Finnish, French, Galician, German, Greek, Hungarian, Italian, Japanese, Korean, Lithuanian, Norwegian, Persian, Polish, Portuguese, Romanian, Russian, Serbian, Sinhala, Slovak, Slovenian, Spanish, Swedish, Turkish, Ukrainian
	<i>Étude(s) de cas publiée(s)</i>	-
	<i>Source(s)</i>	Benedetti 2021 ; COPTR ; Nestor

### 3.1.4 Grille d'analyse des fonctionnalités et des logiciels

Afin de pouvoir prendre connaissance des fonctionnalités détaillées proposées par ces logiciels, les évaluer et les comparer, il était nécessaire de disposer d'un outil méthodologique permettant d'obtenir une vue synthétique et formalisée de l'offre informatique actuelle. Pour cette raison, une grille d'analyse (sous la forme d'un tableau à double entrée) a été établie pour chacune des fonctionnalités sélectionnées (à l'exception de la comparaison de données – nous nous en expliquerons dans la partie qui lui est consacrée). Nous détaillons ci-dessous la méthodologie employée pour l'établissement d'une telle grille.

#### 3.1.4.1 Décomposition des micro-fonctionnalités / informations proposées

Pour chacune des fonctionnalités et des outils sélectionnés, nous avons créé un tableau analytique permettant de décomposer chaque fonctionnalité en une suite plus détaillée de micro-fonctionnalités offertes par les logiciels testés. En ce qui concerne les fonctionnalités dites « actives », il s'agit plus précisément d'une liste de micro-activités permettant le plein accomplissement de la tâche à accomplir (à titre d'exemple, si le dédoublement est une fonctionnalité principale de type « actif », les micro-activités sont notamment l'identification des éléments redondants, les manières de sélectionner les éléments sur lesquels on désire intervenir et les différents types de traitement qu'on leur réserve, allant du déplacement de l'élément à sa suppression définitive). Dans le cas des fonctionnalités dites « descriptives », la forme détaillée de la fonctionnalité consiste en une liste détaillée des informations fournies et des visualisations proposées.

La décomposition fine des fonctionnalités a été menée tout au long de cette recherche, grâce à la lecture des manuels d'utilisation des logiciels et au gré des tests effectués. Un logiciel pouvait effectivement faire apparaître de nouvelles micro-tâches ou pouvait fournir des informations et des visualisations parfois inédites. À l'inverse, il arrivait qu'un logiciel ne présente aucune micro-tâche ou informations totalement inédites, mais l'essai de l'outil pouvait nous amener à préciser la terminologie d'une micro-fonctionnalité ou au contraire à atteindre un niveau d'abstraction plus général, afin d'inclure plusieurs micro-fonctionnalités (très) similaires. Ce travail de conceptualisation / abstraction des micro-fonctionnalités a donc été réalisé de manière *itérative*, l'intitulé et le champ d'action de chaque micro-tâches étant réévalués au fur et à mesure des tests.

#### 3.1.4.2 Attribution des micro-fonctionnalités / informations aux outils testés

Ce tableau à double entrée (les outils par colonne et les micro-fonctionnalités par ligne) a été rempli au fur et à mesure des tests : une valeur positive (1) est insérée dans le tableau lorsque l'outil propose la micro-fonctionnalité ou donne l'information en question ; une valeur nulle (0) est intégrée lorsque le logiciel ne permet pas sa réalisation ou ne livre pas cette information. On veillera à ce que la micro-fonctionnalité et l'information soient proposées de manière évidente et explicite par le logiciel, qu'elles soient facilement accessibles et de qualité (directement compréhensibles, exactes, etc.). En outre, on ne prendra en considération que les micro-fonctionnalités et informations disponibles via l'interface graphique (*GUI*) de l'outil – ce qui exclut, pour l'attribution des valeurs positives et négatives des tableaux analytiques, les informations données en tête des exports des résultats de recherche ou d'extraction de métadonnées, ainsi que les rapports d'audit téléchargeables séparément. Enfin, les micro-fonctionnalités non directement liées à l'évaluation et au tri archivistiques ne sont pas prises en considération, de même que les fonctionnalités les plus triviales et unanimement partagées (on pense par exemple à l'ouverture du répertoire dans un

gestionnaire de fichiers externe de type *Windows Explorer* ou à l'ouverture du fichier dans une application dédiée).

Nous avons également adopté une démarche *itérative* pour l'attribution des valeurs binaires (0 ou 1) traduisant la réalisation d'une micro-fonctionnalité par un logiciel donné : en cas d'apparition d'une nouvelle micro-fonctionnalité via le test d'un logiciel, il était nécessaire de reprendre les essais effectués antérieurement pour contrôler que les logiciels déjà testés n'offraient pas cette possibilité de traitement (que cette micro-fonctionnalité ne nous avait donc pas échappé lors du premier test de l'outil).

### **3.1.4.3 Pondération des résultats**

Au terme des essais de logiciels et de la décomposition détaillée des fonctionnalités, nous avons jugé nécessaire d'attribuer un coefficient à chacune des micro-fonctionnalités listées. Si le codage binaire (0 et 1) s'avère très utile pour savoir ce que font précisément chacun des outils testés et offrir ainsi une vue panoramique de l'offre actuelle en termes de développement informatique, il apporte peu de plus-value pour sélectionner le logiciel grâce auquel on procédera à l'évaluation et au tri des données soumises à l'examen – un logiciel peut effectivement proposer un grand nombre de micro-fonctionnalités, sans que ces dernières soient nécessaires à l'accomplissement d'une tâche ou à la prise de connaissance des contenus. À chaque micro-fonctionnalité correspond donc un coefficient témoignant de son importance et de sa pertinence pour atteindre l'objectif fixé. Cette pondération a été réalisée de manière empirique au terme de la recherche, grâce aux enseignements tirés des études de cas menées avec les données de la Cinémathèque. Les coefficients sont les suivants :

1. Pas ou peu utile pour l'évaluation et le tri archivistiques
2. Peut s'avérer pertinent dans certains cas ou pour une analyse détaillée
3. Nécessaire pour l'évaluation et le tri archivistiques

Cette double notation permet ainsi de calculer le nombre total de micro-fonctionnalités proposées par outil (nombre de valeurs positives qui figurent dans la colonne « Codage ») et un score par logiciel (obtenu en additionnant la pondération attribuée à chaque micro-fonctionnalité / information dans la colonne « Score ») en fonction de la disponibilité de micro-fonctionnalités / informations jugées nécessaires ou au moins pertinentes pour l'évaluation et le tri archivistiques. Enfin, la somme de micro-fonctionnalités / informations par pertinence et par outil pour la tâche à effectuer (coefficient de 1 à 3) est indiquée afin de connaître quel outil propose le plus de fonctionnalités nécessaires et / ou pertinentes (car il peut arriver qu'un outil récolte un score élevé par la multiplication de petites fonctionnalités peu utiles, ou du moins pas directement nécessaires).

## 3.2 Objectifs : une grille d'analyse pérenne et une base d'évaluation sûre

Cette méthode poursuit trois objectifs. Il s'agit tout d'abord de mieux comprendre la manière dont une fonctionnalité informatique (active ou descriptive) et une tâche archivistique sont exécutées par les outils disponibles. Comme nous l'avons évoqué précédemment, il n'est pas rare en effet que les comparatifs d'outils et les études de cas se contentent de lister les fonctionnalités offertes par des logiciels, sans entrer dans le détail de leur réalisation (alors même qu'il existe de grandes différences dans la manière d'exécuter une tâche et dans les informations fournies d'un logiciel à l'autre).

Le tableau analytique permet ensuite de comparer les outils et d'avoir une vue synthétique de leurs forces et faiblesses pour l'exécution d'une tâche spécifique.

Enfin, grâce au coefficient attribué à chaque micro-fonctionnalité, il est possible de connaître les actions ou informations les plus pertinentes pour effectuer une évaluation et opérer un tri des données.

Grâce à cette méthodologie, nous souhaitons donc proposer une grille d'analyse pérenne (puisque de nouvelles fonctionnalités et de nouveaux outils peuvent être ajoutés en tout temps, selon un processus reproductible et itératif), mais aussi fournir aux archivistes une base d'évaluation des outils qui soit la plus objective possible de sorte à éclairer et justifier le choix d'un logiciel. De manière plus marginale vis-à-vis des objectifs de la présente étude, cette méthode peut également s'avérer utile pour la rédaction du cahier des charges d'un futur outil à développer en interne ou à acquérir auprès d'un fournisseur tiers.

## 3.3 Résultats

### 3.3.1 Extraction de métadonnées / Récolement

La comparaison entre les différentes extractions de métadonnées proposées par les six logiciels testés doit comprendre plusieurs niveaux : d'une part des critères sur les types d'extraction possible, d'un point de vue général – ce que l'on trouve dans le « Tableau 5 : Tableau analytique « Extraction de métadonnées / Récolement », qualité générale. » On y a comparé notamment les profondeurs possibles de récolement et le mode de sélection des éléments : peut-on extraire les métadonnées d'une branche de l'arborescence seulement ou doit-on obligatoirement prendre en compte tout le fonds d'archives traité par le logiciel ? Peut-on extraire les métadonnées des seuls dossiers et fichiers se trouvant à un niveau précis de profondeur de l'arborescence ? Peut-on extraire uniquement les métadonnées des éléments qui correspondent à des filtres de recherche particuliers ? Les métadonnées des seuls fichiers, ou des dossiers uniquement ? On s'est également intéressé aux formats possibles d'exportation des métadonnées, ainsi qu'à la qualité du récolement (nous n'avons pas comparé les métadonnées une à une, mais seulement la manière dont les informations sont présentées : les colonnes possèdent-elles des titres explicites ? Ou encore : les informations sont-elles isolées dans des colonnes distinctes ?).

Dans un second temps, nous avons analysé plus précisément les extractions proposées et avons réalisé une cartographie (*mapping*) des métadonnées en fonction de l'intitulé que leur attribuent les différents logiciels, mais également en cherchant à comprendre quel(s) intitulé(s) correspond(ent) à quelle(s) métadonnée(s) (car il est fréquent qu'un logiciel utilise des termes spécifiques pour des métadonnées différentes, tandis qu'un autre range sous le même titre

des métadonnées relatives à des éléments distincts – voire que deux logiciels livrent la même information mais dans des colonnes aux intitulés différents). Le tableau « Tableau 3 : Cartographie des métadonnées extraites » fonctionne ainsi selon une double nomenclature : dans les colonnes par logiciel figure le nom original donné par l'outil à l'information délivrée, tandis que l'on trouve par ligne une proposition de « normalisation » des intitulés (nous avons parfois repris la dénomination utilisée par un logiciel quand ce dernier était le seul à proposer une métadonnée ou une information).

Enfin, grâce à ce *mapping* des métadonnées, nous avons pu établir un tableau récapitulatif (« Tableau 4 : Tableau analytique des métadonnées extraites ») pour voir quelles sont les métadonnées extraites et quelles sont les informations livrées par chacun des outils (avec le système de coefficient pour pondérer l'importance et la pertinence de chaque métadonnée et information).

Il va de soi que le logiciel que nous souhaitons recommander à la Cinémathèque devra au minimum proposer les métadonnées les plus utiles pour une évaluation des contenus enregistrés sur un support de données : le tableau récapitulatif prime ainsi sur les deux autres, puisque certaines métadonnées nous paraissent plus essentielles que d'autres – leur absence dans l'export proposé par un outil aurait donc pour conséquence d'éliminer ce dernier de la liste des logiciels à recommander. La priorité donnée à la liste des métadonnées vis-à-vis du *mapping* (« Tableau 3 : Cartographie des métadonnées extraites ») ou de la qualité générale du récolement (« Tableau 5 : Tableau analytique « Extraction de métadonnées / Récolement », qualité générale ») se justifie aussi par le traitement de données que l'on peut appliquer à un fichier « .csv » ou *Excel* via des outils comme *Tableau Prep* ou *OpenRefine*. Si les colonnes ne sont pas explicites, ou que plusieurs informations sont contenues dans la même colonne, sous le même intitulé, ou que l'on désire reformuler des intitulés de métadonnées, il est possible de le faire dans un second temps, et de créer ainsi son « export idéal ».



### 3.3.1.1 Extraction de métadonnées / Récolement : micro-fonctionnalités et informations

Tableau 1 : « Extraction de métadonnées / Récolement », par métadonnées / informations, résumé

Métadonnées / Informations	Coefficient	Nombre d'outils proposant cette métadonnée / information	Nom des outils proposant la métadonnée / l'information
Chemin complet	3	6	Tous les outils
Date de dernière modification	3	6	Tous les outils
Nom	3	6	Tous les outils
Type d'élément (dossier / fichier)	3	6	Tous les outils
Format de fichier	3	5	Tous sauf Karen
Extension	3	4	Tous sauf Pir, WinCatalog
Chemin contenant	3	3	Pir, TSP, WinCatalog
Nombre de fichiers par dossier individuel	3	3	Karen, Pir, TSP
Nombre total de fichiers subordonnés par branche	3	3	Archifiltre, Pir, TSP
Nombre de dossiers subordonnés par dossier individuel	3	2	Karen, Pir
Nombre total de dossiers subordonnés par branche	3	2	Pir, TSP
Profondeur dans l'arborescence	3	2	Archifiltre, TSP
Indicateur de ressource uniforme (URI)	3	1	DROID
Poids de tous les fichiers dans un dossier individuel	3	1	Karen
MD5 Hash	2	6	Tous les outils
Poids d'un fichier	2	6	Tous les outils
Poids d'un dossier compressé	2	5	Tous sauf Archifiltre
Poids total du dossier/répertoire (sous-dossiers et fichiers)	2	4	Tous sauf DROID, Karen
SHA-256 Hash	2	4	Tous sauf Archifiltre, Pir
Attributs	2	3	Karen, Pir, TSP
Date du dernier accès (lu ou modifié)	2	3	Karen, Pir, TSP
Description	2	3	Archifiltre, TSP, WinCatalog
Auteur	2	1	TSP
Identifiant unique PRONOM du format de fichier (PUID)	2	1	DROID
Mime-type	2	1	DROID
Nom sans extension	2	1	Karen
Propriétaire	2	1	TSP
Date de création	1	4	Tous sauf Archifiltre, DROID
Version du format de fichier	1	3	DROID, Karen, TSP
CRC32	1	2	Pir, WinCatalog
Étiquettes	1	2	Archifiltre, WinCatalog
Longueur du chemin complet	1	2	Archifiltre, TSP
SHA-160 Hash	1	2	DROID, Karen
Avis de non concordance d'extension	1	1	DROID
Date de dernière sauvegarde	1	1	TSP
Date de l'extraction	1	1	TSP
Date de première modification	1	1	Archifiltre
File format count	1	1	DROID
Méthode d'identification ("signature, container signature or extens	1	1	DROID
Nom court (Short Name, format 8.3)	1	1	Karen
Nom utilisateur complet	1	1	TSP
Numéro d'identification personnel de chaque élément	1	1	DROID
Numéro d'identification personnel du dossier parent	1	1	DROID
Poids moyen de fichier	1	1	TSP
Proportion du poids de l'élément vis-à-vis du poids du dossier par	1	1	TSP
Redondances	1	1	Archifiltre
SHA-224 Hash	1	1	Karen
SHA-384 Hash	1	1	Karen
SHA-512 Hash	1	1	Karen
Statut de l'analyse	1	1	DROID

Le tableau ci-dessus présente de manière résumée les résultats de notre enquête<sup>6</sup>. Nous avons identifié 14 métadonnées (sur 53, soit un quart environ) qui nous paraissent essentielles pour mener une évaluation et un tri de données massives. Il s'agit effectivement de l'angle d'approche qui a été privilégié, au détriment très certainement d'autres points de vue, plus techniques, ou plus proches des préoccupations liées à la préservation numérique (c'est ainsi que des champs comme le « PUID » ou le « Mime-type » d'un fichier particulier, ainsi que la version du format dans lequel les données sont encodées ne nous paraissent pas absolument essentiels pour mener à bien la tâche d'évaluation, tandis que ces informations sont capitales pour la préservation et l'accessibilité à long terme des données numériques). Ces métadonnées (au premier rang desquelles on compte le « chemin complet » et le « chemin contenant » des éléments, leur nom, leur format, leur extension et leur profondeur au sein de l'arborescence, ou encore des informations sur le nombre d'items qui sont subordonnés à un dossier individuel ou à une branche entière de répertoire) permettront d'étudier plus finement l'arborescence d'un fonds d'archives numériques, de réaliser quelques analyses statistiques (voir notamment les études de cas qui sont développées dans la partie 4) et de conserver une trace de tous les éléments qui se trouvent dans le fonds avant son traitement. Pour cela, les métadonnées sélectionnées concernent aussi bien les fichiers que les dossiers, sur lesquels l'archiviste a besoin d'avoir des informations pour mener une évaluation *macro* (à l'instar des « séries » dans le domaine analogique).

Un second quart des métadonnées proposées (13 sur 53) nous paraissent potentiellement utiles dans le cadre d'une analyse plus poussée : il s'agit alors le plus souvent d'informations plus détaillées relatives à un fichier particulier (comme le « PUID » et le « Mime-type » dont il a été question ci-dessus, mais également les valeurs de hachage – qui sont extrêmement utiles pour le dédoublement via des logiciels dédiés, mais qui sont peu pertinentes dans un tableur contenant des milliers d'entrées pour l'évaluation) ou encore de champs « date » qui comprennent très souvent des imprécisions (le champ « Date de dernière modification » est semble-t-il le plus fiable, raison pour laquelle nous lui avons attribué un coefficient 3 contrairement à la « Date de création » qui varie dès lors que l'on copie ou déplace le fichier).

Les métadonnées qui nous paraissent peu pertinentes (23 sur 50, soit *grosso modo* la moitié) sont principalement des « champs calculés » (comme la proportion du poids d'un élément vis-à-vis du poids du répertoire parent, le poids moyen des fichiers dans un même dossier, la longueur – en termes de caractères – d'un chemin d'accès, ou encore la colonne « Redondances » d'*Archifiltre* qui propose un code binaire « Oui / NON » suite aux calculs des sommes de contrôle) ; des informations propres au fonctionnement du logiciel ou livrant des informations *méta* à propos de l'analyse elle-même (comme « Statut de l'analyse » ou « Avis de non concordance d'extension » proposés par *Droid*) ; et enfin, des valeurs d'autres algorithmes de hachage de données.

---

<sup>6</sup> Certaines métadonnées proposées par les logiciels n'ont pas été retenues parce qu'elles semblaient trop spécifiques à un outil, trop précises ou trop éloignées de notre objectif d'évaluation et de tri archivistiques. Citons entre autres toutes les données relatives au support de stockage et au serveur qui héberge les données (*TSP* et *WinCatalog*) ; les données relatives au « prêt » potentiel des éléments (*WinCatalog*) ; toutes les balises ID3 pour les fichiers son (*WinCatalog*) ; ou encore les informations qui intéressent surtout les gestionnaires de répertoires partagés, dans un organisme ou une entreprise (*TSP* propose par exemple beaucoup d'informations sur le « pourcentage de croissance » d'un répertoire donné, ou sur les autorisations de consultation et de modification des fichiers – autant d'éléments qui touchent plutôt à la *Gestion Électronique des Documents* et au *record management*).

Au terme de cette évaluation, on peut constater qu'une grande partie des métadonnées peu pertinentes sont proposées par un ou deux outils et que cela dépend donc de la spécialisation du logiciel en question (comme *Droid* qui se consacre surtout à l'identification des fichiers et ne donne quasiment aucune information relative aux dossiers par exemple). En revanche, plusieurs métadonnées essentielles sont proposées par tous les outils. Elles coïncident (presque) avec les métadonnées techniques minimales requises par le logiciel (idéal) pour les tâches « Arrangement and Description » que recommande le projet AIMS – pour les fichiers : « Filename », « Original full file path », « MD5 Hash », « SHA-1 Hash », « Files Dates », « File Size » et « File Format » ; pour les répertoires : « File Count » (« The total number of files within a directory »), « Size » (« Size. Total size of all files in a directory, as expressed in kilobytes, megabytes, gigabytes, etc. ») et « Creation dates » (« Creation dates. A range of all files within the directory ») (AIMS Work Group 2012, p. 136-138). L'identification des métadonnées les plus pertinentes, et qui nous ont été les plus utiles pour les études de cas (via notamment les tris opérés sur les colonnes), devrait permettre soit de « créer » son export de métadonnées idéal à partir des outils existants, soit de paramétrer un logiciel *ad hoc* afin qu'il fournisse aux archivistes en charge de l'évaluation les métadonnées nécessaires.

### 3.3.1.2 Extraction de métadonnées / Récolement : *mapping* des métadonnées

Le cartographie (*mapping*) des métadonnées donne un aperçu complet de la manière dont les métadonnées sont extraites et comment sont structurés les récolements. En effet, on peut ainsi très facilement prendre connaissance des intitulés de colonnes des différents tableurs générés par les outils, et constater les chevauchements possibles d'un outil à l'autre et l'importance accordée par chacun d'eux aux différentes informations. Cette cartographie met aussi en évidence la diversité des encodages possibles, avec les difficultés que cela peut engendrer pour comparer les extractions et les logiciels (et décider de l'outil que l'institution patrimoniale doit adopter).

On trouve plusieurs grandes orientations et quelques champs qui posent des problèmes spécifiques. Si l'on voit facilement que certains logiciels ne scannent que les fichiers et proposent beaucoup de champs très « techniques » (comme *Droid*, à propos des formats), d'autres livrent quantité d'informations sur les dossiers (comme *Karen's Directory Printer*), ou offrent beaucoup de « champs calculés » de sorte à faciliter la prise en main et la compréhension du récolement (comme *Archifiltre* ou *TreeSize*, avec des proportions ou des informations sur la longueur des chemins – autant de valeurs que l'on pourrait facilement générer via des formules dans *Excel* si le besoin s'en faisait sentir). Au-delà de ces orientations (fichiers / dossiers, métadonnées originales / champs calculés, public spécialisé / grand public), on constate que les logiciels n'adoptent pas la même granularité quant aux intitulés de colonnes dans les récolements qu'ils proposent. *Karen's Directory Printer* se distingue tout particulièrement par la grande subdivision des champs proposés (on compte ainsi des champs distincts pour le nom et le chemin contenant d'un fichier ou d'un dossier) tandis que *Pir* regroupe souvent une grande quantité d'informations (nécessaires ou pertinentes pour la plupart) dans un même champ : sous l'intitulé « Size » on trouve par exemple *toutes* les données relatives au nombre *et* au poids des éléments. Pour ce qui concerne la comparaison des champs et des intitulés *inter-logiciels*, l'exemple le plus frappant des difficultés que l'on peut rencontrer est le mot « Type » dont voici quelques « déclinaisons ».

- *Archifiltre* : sous l'intitulé « Type », le logiciel français indique le format du fichier (et « répertoire » pour les dossiers de fichiers) ; la distinction entre les fichiers et les répertoires se trouve alors dans une autre colonne, intitulée opportunément « Fichier/Répertoire » ;
- *Droid* : le champ « Type » contient les trois informations suivantes : « file », « folder » ou « container » tandis que le nom du format est indiqué dans le champ « Format name ».
- *Karen's Directory Printer* : il n'existe pas de champ explicitement intitulé « Type », mais une colonne, sans titre aucun, contient les informations « File » ou « Folder ».
- *Pir* indique dans le même champ (fréquemment disposé en ligne) et sous le seul intitulé « Type », le type d'éléments (dossier/fichier) et le format du fichier.

Au vu des grandes différences qui peuvent exister entre les logiciels (pour l'orientation choisie, la granularité des champs ou l'intitulé des colonnes) dans l'extraction et l'export des métadonnées proposées, cette cartographie peut donc s'avérer un outil très utile si l'on désire faire une réconciliation entre plusieurs exports. Il n'existe en effet pas un *seul* outil qui réponde à tous les besoins des archivistes pour prendre connaissance des contenus stockés sur un support de données via une extraction de métadonnées : il sera alors nécessaire de « compiler » les récolements des différents outils, c'est-à-dire de faire une « réconciliation » des données (via une clé relationnelle) en sélectionnant, pour chaque métadonnée, quel est l'outil qui en propose la meilleure représentation.

### 3.3.1.3 Extraction de métadonnées / Récolement : outils

Les deux logiciels qui sortent globalement en tête de cette étude des métadonnées proposées et du type d'extraction qu'il est possible de réaliser sont : *TreeSize Professional* et *Karen's Directory Printer* (en particulier pour les métadonnées proposées, plus encore que pour les caractéristiques générales de l'extraction et de l'exportation). Ces deux logiciels se distinguent également dans leur catégorie : *TSP* parmi les logiciels généralistes ; *Karen's Directory Printer* parmi les logiciels spécialisés.

Tableau 2 : « Extraction de métadonnées / Récolement », par outils, résumé

Bilan final	Outils généralistes				Outils spécialisés	
	Archifiltre	DROID	TreeSize Professional	WinCatalog	Karen	Pir
<b>Métadonnées proposées</b>						
Informations données	16	21	29	15	23	17
Score pondéré	36	41	61	33	48	41
Métadonnées pas ou peu utiles (1)	4	8	8	3	7	2
Métadonnées pertinentes (2)	4	6	10	6	7	6
Métadonnées nécessaires (3)	8	7	11	6	9	9
<b>Caractéristiques générales</b>						
Informations données	7	7	18	13	15	12
Score pondéré	19	18	38	31	29	25
Fonctionnalités pas ou peu utiles (1)	0	1	5	2	5	3
Fonctionnalités pertinentes (2)	2	1	6	4	6	5
Métadonnées nécessaires (3)	5	5	7	7	4	4

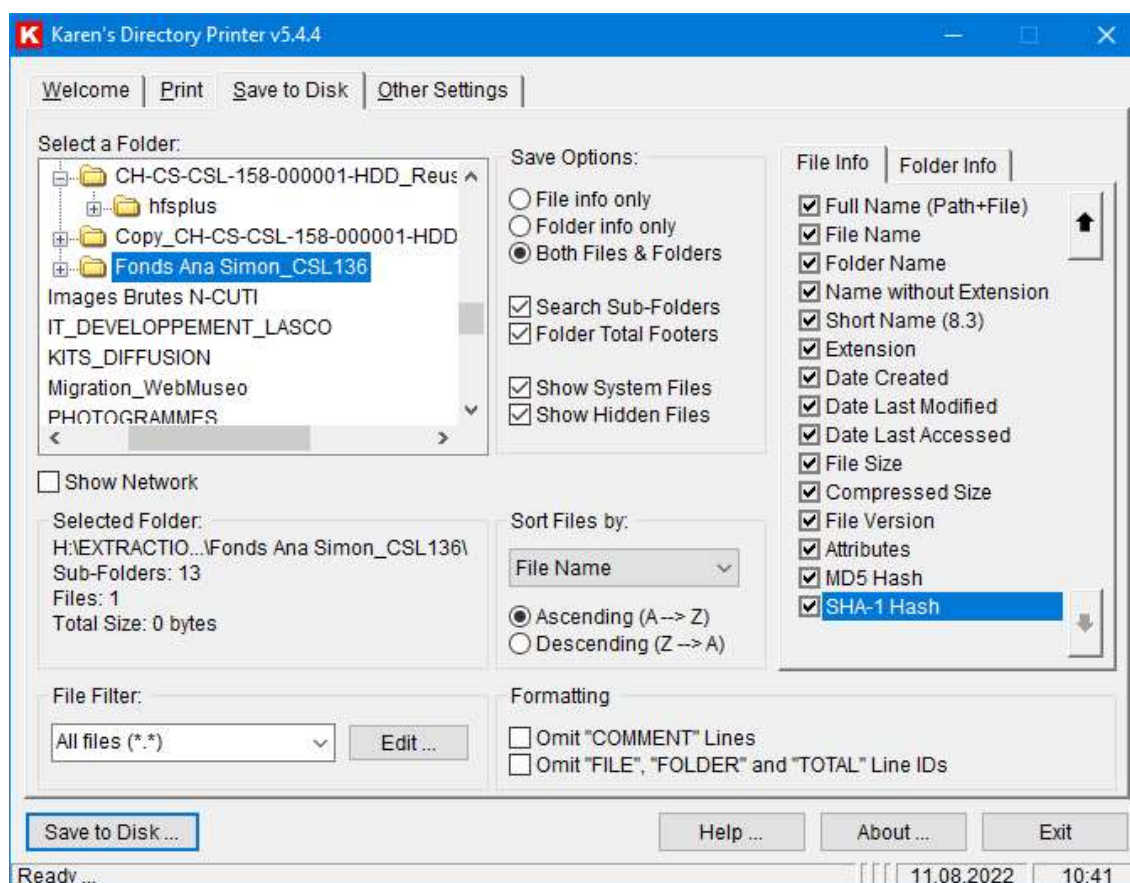
*Archifiltre*, *Droid*, et *TreeSize* sont présentés plus en détails par la suite, en lien avec l'étude de l'arborescence et du dédoublonnage, raison pour laquelle nous ne nous attardons pas sur ces outils dans cette partie consacrée à l'extraction, au profit de la présentation de l'outil proposé par *Karen's Power Tools* (du nom de sa fondatrice, Karen Kenworthy) : *Karen's Directory Printer*. Cet outil fait effectivement partie d'une suite de logiciels bureautiques

simples d'utilisation et proposés gratuitement par des développeurs privés (« Karen's Power Tools are utility programs that make life with Microsoft Windows a lot easier », disent d'ailleurs les développeurs sur la page d'accueil du site (Karenware 2022)). Si la première version a été lancée en 1997 déjà, le logiciel est encore récemment mis à jour (la dernière version date de mai 2020) et semble très bien diffusé au sein de la communauté archivistique. Des projets aussi importants et de grande envergure que *Paradigm* et *AIMS* citent ainsi cet outil parmi les (rares) logiciels qu'ils testent ou recommandent :

*« Paradigm experimented with tools such as DirPrinting and Karen's Directory Printer, which can be used to generate a complete list of all folders, subfolders and files in a directory providing the archivist with an overview of the accession » (Paradigm Project 2007, p. 45)*

*« It is possible to create file and folder information at the same time, but having two separate manifests makes using the data in further tasks easier. Indeed this is one factor that has meant we have decided to keep using this software despite similar functionality being offered by FTK Imager. Although its use involves another piece of software in our workflow we felt the tool was simple and easy to use and feel confident in suggesting its use by depositors who may wish to create a list of files that they intend to transfer. » (AIMS Work Group 2012, p. 128)*

Figure 8 : Interface graphique principale de *Karen's Directory Printer*, Fonds Simon



*Karen's Directory Printer* offre effectivement la possibilité de créer des récolements sélectifs pour les dossiers et fichiers, ou pour les seuls fichiers / dossiers (« File info only » ou « Folder info only »). Si cette fonctionnalité peut paraître relativement triviale (elle n'est en tout cas pas difficile à intégrer dans un logiciel, d'un point de vue technique), elle n'est pourtant pas proposée par tous les logiciels testés (seuls *Karen's Directory Printer* et *Pir* offrent par exemple

la possibilité de créer des listes des seuls fichiers). En outre, *Karen's Directory Printer* extrait des métadonnées particulièrement intéressantes pour l'évaluation d'un fonds d'archives numériques : le champ « Name without Extension » et le nombre de dossiers et de fichiers qui se trouvent dans un répertoire *individuel*. Le nom sans extension permet par exemple de filtrer les résultats pour identifier tous les fichiers qui possèdent le même intitulé mais qui sont encodés dans des formats différents (un exemple en est donné dans la partie « 4.5.2 Comparer les métadonnées des fichiers » avec des éléments possédant le même nom avec des extensions « .doc » et « .pdf »). La manière de compter les fichiers et dossiers est enfin une spécialité de *Karen's Directory Printer* qui le rend si utile d'après nos études de cas : la plupart des logiciels livre le nombre *total* des fichiers et dossiers que compte un répertoire, c'est-à-dire *tous* les éléments qui se trouvent à un niveau de profondeur subordonné, du niveau subordonné *n-1* jusqu'au dernier niveau de l'arborescence. À l'inverse, *Karen's Directory Printer* indique le nombre de sous-dossiers et de fichiers que contient un dossier *individuel* au niveau *n-1*, c'est-à-dire au premier niveau de subordination. À titre d'exemple, le dossier sommital du disque dur de Francis Reusser, de premier niveau (« hfsplus », *n*), contient, au niveau qui lui est directement subordonné (*n-1*), 17 sous-dossiers et 15 fichiers (le logiciel ne retourne donc pas le nombre total de dossiers et fichiers que contient *en tout* ce dossier de niveau 1, qui se monte à plusieurs dizaines de milliers). On voit ci-dessous les valeurs affichées pour le premier niveau, ainsi que pour quelques-uns des dossiers de niveau 2 (« A CLASSER », « aaCINEATELIER », « aaREUSSER », etc.) :

Figure 9 : Récolement proposé par *Karen's Directory Printer*, Fonds Reusser

COMMENT Karen's Directory Printer v5.4.4

COMMENT AG 1997, 1999-2002, 2004-2005, 2007-2008 by Karen Kemsworthy, AG 2018-2020 by Joe Winett DBA KarenWare.com - All Rights Reserved

COMMENT http://www.karenware.com/

COMMENT

COMMENT Lines beginning with "COMMENT" (like this one) contain comments or explanations.

COMMENT They do not contain any information about particular files or folders.

COMMENT

COMMENT Lines beginning with "FOLDER" contain information about a single folder.

COMMENT Here's the information found in those lines:

COMMENT Full Name (Path+Folder) <TAB> Folder Name <TAB> Short Name (8.3) <TAB> Parent Folder <TAB> Number of Sub-Folders <TAB> Number of Files <TAB> Folder Size <TAB> Compressed Size <TAB> Attributes <TAB> Date Created <TAB> Date Last Modified <TAB> Date Last Accessed

COMMENT

COMMENT Computer: PCL-P2-0110

COMMENT User: denis.bussard

COMMENT Prepared: 14:35 07/07/2022

COMMENT Folder: H:\EXTRACTION\TM\_Bussard\_Denis\CH-CS-CSL-158-000001-HDD\_Reusser\_LaCie\_Quadra\hfsplus\

COMMENT Include Sub-Folders? Yes

COMMENT

FOLDER	CH-CS-CSL-158-000001-HDD_Reusser_LaCie_Quadra\hfsplus\	hfsplus	hfsplus	CH-CS-CSL-158-000001-HDD_Reusser_LaCie_Quadra\	17	15	180'490'983	180'490'983	18.02.2022 11:35	15.01.2017 13:42	07.07.2022 13:25
FOLDER	CH-CS-CSL-158-000001-HDD_Reusser_LaCie_Quadra\hfsplus\A CLASSER	A CLASSER	ACLAS	CH-CS-CSL-158-000001-HDD_Reusser_LaCie_Quadra\hfsplus\	9	50	747'127'84	747'127'84	18.02.2022 11:39	13.07.2014 12:12	07.07.2022 13:30
FOLDER	CH-CS-CSL-158-000001-HDD_Reusser_LaCie_Quadra\hfsplus\aaCINEATELIER	aaCINEATELIER	AACH-905	CH-CS-CSL-158-000001-HDD_Reusser_LaCie_Quadra\hfsplus\	44	1	307'24	307'24	18.02.2022 13:05	14.01.2013 20:27	07.07.2022 13:32
FOLDER	CH-CS-CSL-158-000001-HDD_Reusser_LaCie_Quadra\hfsplus\aaREUSSER	aaREUSSER	AARE-1M	CH-CS-CSL-158-000001-HDD_Reusser_LaCie_Quadra\hfsplus\	20	2	104'872'964	104'872'964	18.02.2022 13:52	22.03.2013 18:39	07.07.2022 14:22
FOLDER	CH-CS-CSL-158-000001-HDD_Reusser_LaCie_Quadra\hfsplus\ADRESSSES \			CH-CS-CSL-158-000001-HDD_Reusser_LaCie_Quadra\hfsplus\	6	1	12'292	12'292	invalid	invalid	invalid
FOLDER	CH-CS-CSL-158-000001-HDD_Reusser_LaCie_Quadra\hfsplus\EMMANUELLE	EMMANUELLE	EMMA-QM	CH-CS-CSL-158-000001-HDD_Reusser_LaCie_Quadra\hfsplus\	6	2	408'570	408'570	18.02.2022 11:41	14.01.2013 20:25	07.07.2022 14:02
FOLDER	CH-CS-CSL-158-000001-HDD_Reusser_LaCie_Quadra\hfsplus\FILMS	FILMS	FILMS	CH-CS-CSL-158-000001-HDD_Reusser_LaCie_Quadra\hfsplus\	3	2	8'689'043	8'689'043	18.02.2022 11:43	15.01.2017 15:39	07.07.2022 14:24
FOLDER	CH-CS-CSL-158-000001-HDD_Reusser_LaCie_Quadra\hfsplus\JEAN NEW	JEAN NEW	JEAN NEW	CH-CS-CSL-158-000001-HDD_Reusser_LaCie_Quadra\hfsplus\	3	6	1'509'526'432	1'509'526'432	18.02.2022 12:04	28.11.2012 17:26	07.07.2022 14:34
FOLDER	CH-CS-CSL-158-000001-HDD_Reusser_LaCie_Quadra\hfsplus\UTILE	UTILE	UTILE	CH-CS-CSL-158-000001-HDD_Reusser_LaCie_Quadra\hfsplus\	9	3	1'059'961'950	1'059'961'950	18.02.2022 12:05	13.07.2014 12:12	07.07.2022 14:24

On peut seulement regretter que *Karen's Directory Printer* ne fournisse pas le niveau de profondeur des éléments (cette valeur peut cependant être calculée dans *Excel* via une formule qui dénombre les barres obliques inversées – *backslash* – qui séparent les répertoires par niveau) et, dans le même ordre d'idée, qu'il n'offre pas la possibilité de récoiler uniquement les éléments de niveau 1, 2, 3 et ainsi de suite, comme le proposent *TreeSize* ou *Pir* (une fonctionnalité qui permet de connaître plus spécifiquement quels sont les dossiers et fichiers par profondeur et de faire apparaître un potentiel plan de classification thématique, chronologique, etc.). Enfin, dernières réserves, le logiciel ne fournit pas un export très « propre » puisque les intitulés de colonne de figurent pas de manière explicite (ils sont listés dans les commentaires généraux du récolement comme on peut le voir ci-dessus), et les données relatives aux fichiers et aux dossiers (lorsque l'on exporte les métadonnées des deux types d'éléments) sont insérées à la suite et, puisque les colonnes ne correspondent pas exactement, on obtient quelques décalages troublants – ces défauts peuvent cependant être corrigés moyennant quelques étapes de nettoyage des données dans un tableur.

#### **3.3.1.4 Extraction de métadonnées / Récolement : conclusion**

Cette enquête a permis de mettre en avant l'embrouillamini que constitue la comparaison de l'extraction de métadonnées : les champs ne sont pas toujours tout à fait les mêmes, les informations peuvent être regroupées dans un même champ, la granularité, la précision et la nomenclature diffèrent, etc. Obtenir un bon récolement n'est donc pas chose aisée, et il est nécessaire de choisir les métadonnées nécessaires à l'évaluation, de sélectionner les bons outils et d'avoir quelques compétences en traitement de données pour aboutir à un résultat satisfaisant et prêt à l'emploi. Car c'est bien tout l'enjeu : si l'établissement d'un récolement de qualité constitue la première étape de l'analyse – puisqu'il s'agit d'un instantané fidèle des contenus du support de stockage tel qu'il a été remis par le producteur des documents ou ses ayants droit –, ce n'est « que » le début du travail. Aussi difficile à obtenir soit-il, un bon récolement n'est pas une fin en soi quand on désire évaluer et trier des données massives ; il est un prérequis nécessaire pour les étapes de travail à venir (nous donnons des exemples d'utilisation des récolements dans les études de cas).

Tableau 3 : Cartographie des métadonnées extraites

Mapping des métadonnées	Domaines	Nom normalisé	Coefficient	Outils généralistes				Outils spécialisés	
				Archifiltre	DROID	TreeSize Professionnal	WinCatalog	Karen's Directory Printer	Pir
				Noms originaux des champs					
Métadonnées	Identification et arborescence	Numéro d'identification personnel de chaque élément	1	x	ID	x	x	x	x
		Numéro d'identification personnel du dossier parent	1	x	Parent ID	x	x	x	x
		Indicateur de ressource uniforme (URI)	3	x	Unique resource indicator (URI)	x	x	x	x
		Chemin contenant	3	x	x	Chemin contenant	Chemin du catalogue	Parent Folder Folder Name	x
		Chemin complet	3	Chemin d'accès	File path	Chemin complet	Chemin du fichier	Full Name (Path + File)	Path
		Longueur du chemin complet	1	Longueur du chemin d'accès	x	Longueur du chemin	x	x	x
		Profondeur dans l'arborescence	3	Profondeur	x	Niveau de répertoire	x	x	x
						Niveau de répertoire (relatif)			
	Nom de l'élément	Nom	3	Nom de l'élément	Filename	Nom	Nom de fichier	File Name Folder Name	Name
		Nom court (Short Name, format 8.3)	1	x	x	x	x	Short Name (8.3)	x
		Nom sans extension	2	x	x	x	x	Name without Extension	x
	Volumétrie - Nombre d'éléments	Nombre total de fichiers subordonnés par branche	3	Nombre de fichiers	x	Fichiers	x	x	
		Nombre de fichiers par dossier individuel	3	x	x	Dossiers	x	Number of Files	Size
		Nombre total de dossiers subordonnés par branche	3	x	x	x	x	Number of Sub-Folders	
		Nombre de dossiers subordonnés par dossier individuel	3	x	x	x	x	Folder Size	x
	Volumétrie - Poids des éléments	Poids de tous les fichiers dans un dossier individuel	3	x	x	x	x	x	
		Poids total du dossier/répertoire (sous-dossiers et fichiers)	2	Poids (en octet)	x			File size	Size
		Poids d'un fichier	2		File size	Taille	Taille	File size	
		Poids d'un dossier compressé	2	x				Compressed Size	
		Poids moyen de fichier	1	x	x	Taille de fichier moy	x	x	x
	Date des éléments	Proportion du poids de l'élément vis-à-vis du poids du dossier parent	1	x	x	% du parent [% of Parent]	x	x	x
		Date de création	1	x	x	Date de création	Créé le	Date Created	Created
		Date de première modification	1	Date de première modification	x	x	x	x	x
		Date de dernière modification	3	Date de dernière modification	Last modified date	Dernière modification ("Last Modified")	Modifié le	Date Last Modified	Changed [Modified]
		Date de dernière sauvegarde	1	x	x	Date de dernière modification ("Last Save Date")	x	x	x
	Identification de l'élément	Date du dernier accès (lu ou modifié)	2	x	x	Dernier accès	x	Date Last Accessed	Accessed
		Date de l'extraction	1	x	x	Date actuelle ("Current Date")	x	x	x
		Type d'élément (dossier / fichier)	3	Fichier / Répertoire	Type (file, folder or container)	Type	Type	(Sans titre : "File", "Folder")	Type
		Extension	3	Extension	File extension	Extension	x	Extension	x
		Avis de non concordance d'extension	1	x	Extension mismatch warning	x	x	x	x
		Format de fichier	3	Type	Format name	Type	Type	x	Type
		Versión du format de fichier	1	x	File format version	Versión du fichier	x	File Version	x
		Attributs	2	x	x	Attributs	x	Attributes	Attributes
		Méthode d'identification ("signature, container signature or extension")	1	x	Identification method (signature, container signature or extension)	x	x	x	x
		Statut de l'analyse	1	x	Status	x	x	x	x
	Empreintes numériques	Identifiant unique PRONOM du format de fichier (PUID)	2	x	PUID	x	x	x	x
		File format count	1	x	File format count	x	x	x	x
		Mime-type	2	x	Mime-type	x	x	x	x
		MD5 Hash	2	Empreintes [MD5]		Somme de contrôle MD5	MD5 / SHA256 / Blake2	MD5 Hash	MD5
		SHA-256 Hash	2	x	Hash	Somme de contrôle SHA256		SHA-256 Hash	x
		SHA-160 Hash	1	x		x	x	SHA-1 Hash	x
		SHA-224 Hash	1	x	x	x	x	SHA-224 Hash	x
Ajouts / Enrichissement	Utilisateur	SHA-384 Hash	1	x	x	x	x	SHA-384 Hash	x
		SHA-512 Hash	1	x	x	x	x	SHA-512 Hash	x
		CRC32	1	x	x	x	CRC32		CRC32
	Description	Redondances	1	Redondances [OUI / NON]	x	x	x	x	x
		Auteur	2	x	x	Auteur	x	x	x
		Propriétaire	2	x	x	Propriétaire	x	x	x
	Ajouts / Enrichissement	Nom utilisateur complet	1	x	x	Nom utilisateur complet	x	x	x
		Description	2	Description	x	Description	Description	x	x
	Etiquettes	Etiquettes	1	Etiquette (tag)	x	x	tags associated	x	x



Tableau 4 : Tableau analytique des métadonnées extraites

Fonctionnalité	Domaines	Nom normalisé	Coefficient	Outils généralistes								Outils spécialisés				Outils offrant cette fonctionnalité
				Archifiltre		DROID		TreeSize Professionnal		WinCatalog		Karen's Directory Printer		Pir		
				Codage	Score	Codage	Score	Codage	Score	Codage	Score	Codage	Score	Codage	Score	
Extraction de métadonnées / Récolement	Identification et arborescence	Numéro d'identification personnel de chaque élément	1	0	0	1	1	0	0	0	0	0	0	0	0	1
		Numéro d'identification personnel du dossier parent	1	0	0	1	1	0	0	0	0	0	0	0	0	1
		Indicateur de ressource uniforme (URI)	3	0	0	1	3	0	0	0	0	0	0	0	0	1
		Chemin contenant	3	0	0	0	0	1	3	1	3	1	3	0	0	3
		Chemin complet	3	1	3	1	3	1	3	1	3	1	3	1	3	6
		Longueur du chemin complet	1	1	1	0	0	1	1	0	0	0	0	0	0	2
		Profondeur dans l'arborescence	3	1	3	0	0	1	3	0	0	0	0	0	0	2
	Nom de l'élément	Nom	3	1	3	1	3	1	3	1	3	1	3	1	3	6
		Nom court (Short Name, format 8.3)	1	0	0	0	0	0	0	0	0	1	1	0	0	1
		Nom sans extension	2	0	0	0	0	0	0	0	0	1	2	0	0	1
	Volumétrie - Nombre d'éléments	Nombre total de fichiers subordonnés par branche	3	1	3	0	0	1	3	0	0	0	0	1	3	3
		Nombre de fichiers par dossier individuel	3	0	0	0	0	1	3	0	0	1	3	1	3	3
		Nombre total de dossiers subordonnés par branche	3	0	0	0	0	1	3	0	0	0	1	3	3	2
		Nombre de dossiers subordonnés par dossier individuel	3	0	0	0	0	0	0	0	0	1	3	1	3	2
	Volumétrie - Poids des éléments	Poids de tous les fichiers dans un dossier individuel	3	0	0	0	0	0	0	0	0	1	3	0	0	1
		Poids total du dossier/répertoire (sous-dossiers et fichiers)	2	1	2	0	0	1	2	1	2	0	0	1	2	4
		Poids d'un fichier	2	1	2	1	2	1	2	1	2	1	2	1	2	6
		Poids d'un dossier compressé	2	0	0	1	2	1	2	1	2	1	2	1	2	5
		Poids moyen de fichier	1	0	0	0	0	1	1	0	0	0	0	0	0	1
		Proportion du poids de l'élément vis-à-vis du poids du dossier parent	1	0	0	0	0	1	1	0	0	0	0	0	0	1
	Date des éléments	Date de création	1	0	0	0	0	1	1	1	1	1	1	1	1	4
		Date de première modification	1	1	1	0	0	0	0	0	0	0	0	0	0	1
		Date de dernière modification	3	1	3	1	3	1	3	1	3	1	3	1	3	6
		Date de dernière sauvegarde	1	0	0	0	0	1	1	0	0	0	0	0	0	1
		Date du dernier accès (lu ou modifié)	2	0	0	0	0	1	2	0	0	1	2	1	2	3
		Date de l'extraction	1	0	0	0	0	1	1	0	0	0	0	0	0	1
	Identification de l'élément	Type d'élément (dossier / fichier)	3	1	3	1	3	1	3	1	3	1	3	1	3	6
		Extension	3	1	3	1	3	1	3	0	0	1	3	0	0	4
		Avis de non concordance d'extension	1	0	0	1	1	0	0	0	0	0	0	0	0	1
		Format de fichier	3	1	3	1	3	1	3	1	3	0	0	1	3	5
		Versión du format de fichier	1	0	0	1	1	1	1	0	0	1	1	0	0	3
		Attributs	2	0	0	0	0	1	2	0	0	1	2	1	2	3
		Méthode d'identification ("signature, container signature or extension")	1	0	0	1	1	0	0	0	0	0	0	0	0	1
		Statut de l'analyse	1	0	0	1	1	0	0	0	0	0	0	0	0	1
		Identifiant unique PRONOM du format de fichier (PUID)	2	0	0	1	2	0	0	0	0	0	0	0	0	1
		File format count	1	0	0	1	1	0	0	0	0	0	0	0	0	1
		Mime-type	2	0	0	1	2	0	0	0	0	0	0	0	0	1
	Empreintes numériques	MD5 Hash	2	1	2	1	2	1	2	1	2	1	2	1	2	6
		SHA-256 Hash	2	0	0	1	2	1	2	1	2	1	2	0	0	4
		SHA-160 Hash	1	0	0	1	1	0	0	0	0	1	1	0	0	2
		SHA-224 Hash	1	0	0	0	0	0	0	0	0	1	1	0	0	1
		SHA-384 Hash	1	0	0	0	0	0	0	0	0	1	1	0	0	1
		SHA-512 Hash	1	0	0	0	0	0	0	0	0	1	1	0	0	1
		CRC32	1	0	0	0	0	0	0	1	1	0	0	1	1	2
		Redondances	1	1	1	0	0	0	0	0	0	0	0	0	0	1
	Utilisateur	Auteur	2	0	0	0	0	1	2	0	0	0	0	0	0	1
		Propriétaire	2	0	0	0	0	1	2	0	0	0	0	0	0	1
		Nom utilisateur complet	1	0	0	0	0	1	1	0	0	0	0	0	0	1
	Ajouts / Enrichissement	Description	2	1	2	0	0	1	2	1	2	0	0	0	0	3
		Étiquettes	1	1	1	0	0	0	0	1	1	0	0	0	0	2
Bilan final	Nombre absolu de métadonnées et score pondéré			16	36	21	41	29	61	15	33	23	48	17	41	
	Nombre de métadonnées pas ou peu utiles				4		8		8		3		7		2	
	Nombre de métadonnées pertinentes				4		6		10		6		7		6	
	Nombre de métadonnées nécessaires				8		7		11		6		9		9	

Tableau 5 : Tableau analytique « Extraction de métadonnées / Récolement », qualité générale

Fonctionnalité	Domaines	Micro-fonctionnalités	Coefficient	Outils généralistes								Outils spécialisés				Outils offrant cette fonctionnalité
				Archifiltre		DROID		TreeSize Professional		WinCatalog		Karen's Directory Printer		Pir		
				Codage	Score	Codage	Score	Codage	Score	Codage	Score	Codage	Score	Codage	Score	
Extraction de métadonnées / Récolement	Sélection de la profondeur du récolement	Par arborescence complète (toutes les données)	3	1	3	1	3	1	3	1	3	1	3	1	3	6
		Par filtres de recherche (exportation des résultats)	3	1	3	1	3	1	3	1	3	0	0	0	0	4
		Par éléments (dossiers ou fichiers)	2	0	0	0	0	1	2	1	2	1	2	1	2	4
		Par sélection de répertoires	2	0	0	0	0	0	0	1	2	1	2	1	2	3
		Par niveau de profondeur de l'arborescence	2	0	0	0	0	1	2	0	0	0	0	1	2	2
		Par format de fichiers (extensions ou groupes)	2	0	0	0	0	0	0	0	0	1	2	0	0	1
		Par poids des éléments	2	0	0	0	0	1	2	0	0	0	0	0	0	1
		Par éléments visibles dans l'interface graphique utilisateur	1	0	0	0	0	1	1	0	0	0	0	0	0	1
	Éléments extraits	Dossiers et fichiers	3	1	3	1	3	1	3	1	3	1	3	1	3	6
		Dossiers	2	0	0	0	0	1	2	1	2	1	2	1	2	4
		Fichiers	2	0	0	0	0	0	0	0	0	1	2	1	2	2
		Dossiers compressés	1	0	0	1	1	0	0	1	1	0	0	0	0	2
	Formats d'exportation	Export CSV	3	1	3	1	3	1	3	1	3	0	0	1	3	5
		Export XML	3	0	0	0	0	1	3	1	3	0	0	0	0	2
		Export Excel	2	1	2	0	0	1	2	0	0	0	0	0	0	2
		Export TSV	2	0	0	0	0	0	0	0	0	1	2	0	0	1
		Export PDF	1	0	0	0	0	1	1	0	0	1	1	1	1	3
		Export TXT	1	0	0	0	0	1	1	0	0	1	1	1	1	3
		Export HTML	1	0	0	0	0	1	1	1	1	0	0	0	0	2
		Export RTF	1	0	0	0	0	0	0	0	0	0	0	1	1	1
		Export "Print-to-disk" (.prn)	1	0	0	0	0	0	0	0	0	1	1	0	0	1
		Export "Folder information" (.dir)	1	0	0	0	0	0	0	0	0	1	1	0	0	1
	Disposition des informations	Information isolée par colonne	3	1	3	1	3	1	3	1	3	1	3	0	0	5
		Intitulé explicite des colonnes	2	1	2	1	2	1	2	1	2	0	0	0	0	4
	Paramétrabilité	Choix des métadonnées	3	0	0	0	0	1	3	1	3	1	3	1	3	4
		Ordre des métadonnées	1	0	0	0	0	1	1	0	0	1	1	0	0	2
	Bilan final	Nombre absolu de micro-fonctionnalités et score pondéré		7	19	7	18	18	38	13	31	15	29	12	25	
		Nombre de fonctionnalités pas ou peu utiles			0		1		5		2		5		3	
		Nombre de fonctionnalités pertinentes			2		1		6		4		6		5	
		Nombre de fonctionnalités nécessaires			5		5		7		7		4		4	

### 3.3.2 Arborescence et volumétrie

#### 3.3.2.1 Arborescence et volumétrie : micro-fonctionnalités et informations

Grâce aux tests de cinq logiciels menés avec les données des Fonds Reusser et Simon, 40 micro-fonctionnalités ou informations ont pu être isolées (Tableau 8 : Tableau analytique « Arborescence et volumétrie »). Ces dernières sont regroupées dans huit domaines distincts qui recoupent en particulier les visualisations proposées de l'arborescence, de la volumétrie ou du type d'extensions (sous forme de graphiques ou via l'interface utilisateur du logiciel) ainsi que les informations quantitatives relatives à l'ensemble des données ou à une partie seulement de l'arborescence.

Parmi ces informations, 14 (soit 35 %) nous paraissent nécessaires à l'évaluation ; 17 (42,5 %) sont jugées pertinentes dans le cadre d'une analyse plus détaillée ; et 9 (22,5 %) sont pas ou peu pertinentes pour le tri archivistique des supports contenant un nombre très conséquent de données.

Comme on peut le constater dans le « Tableau 6 : « Arborescence et volumétrie », par informations et micro-fonctionnalités, résumé », les informations prioritaires (coefficient 3) sont globalement proposées par la majorité des logiciels testés, tandis que les informations pas ou peu pertinentes sont délivrées par un ou deux logiciels au maximum. Il apparaît donc qu'il existe un consensus sur les micro-fonctionnalités que doit posséder un logiciel pour être largement diffusé et être référencé dans les listes d'outils tenues à jour par des archivistes professionnels. Ces informations concernent en premier lieu les statistiques relatives à l'ensemble des éléments contenus sur un support de données, c'est-à-dire au nombre et au poids total des dossiers et fichiers dans l'arborescence complète ou dans une branche particulière. De même, tous les outils proposent des informations à propos des formats de fichiers, que ce soit sous la forme d'extensions seules, ou de types de fichiers regroupant plusieurs extensions différentes (fichiers images, sons, vidéos, etc.). Nous privilégions toutefois l'information par extension plutôt que par type pour les deux raisons suivantes : les regroupements ne sont pas toujours paramétrables et les extensions contenues dans un groupe ne sont pas toujours listées de manière explicite (c'est par exemple le cas d'*Archifiltre* qui donne la liste des extensions dans le seul rapport d'audit et qui ne permet aucune configuration manuelle des types) ; et les directives relatives à la préservation numérique s'appliquent aux formats (quels sont les formats privilégiés pour la conservation et l'accessibilité à long terme) et non aux types de fichiers. Enfin, la quasi-totalité des logiciels (à l'exception d'*Archifiltre*) proposent une navigation au sein de l'arborescence via un gestionnaire de fichiers<sup>7</sup>, de type *Windows Explorer* (PC) ou *Finder* (Mac) – ce qui facilite grandement la prise en main du logiciel puisque cela correspond à l'environnement bureautique avec lequel nous interagissons sur nos ordinateurs personnels et professionnels. De manière générale, les informations qui se rapportent aux niveaux de l'arborescence les plus élevés (le support lui-même et les branches particulières de l'arborescence) sont

---

<sup>7</sup> Compris comme un « Software used to organize files on a storage device (hard drive, SSD, flash drive). The file manager displays the file/folder hierarchy, and it provides functions to create, copy, move, rename and delete folders as well as copy, move, rename and delete files » (PCMag 2022). Permettant de gérer l'espace de stockage et de naviguer dans le système de fichiers, un gestionnaire prend le plus souvent la forme d'une structure hiérarchique de dossiers et fichiers (le gestionnaire de fichiers est appelé *Windows Explorer* ou simplement *Explorer* depuis Windows 95 pour les utilisateurs de PC, et fait partie de la suite *Finder* pour les utilisateurs d'ordinateurs Macintosh).

privilégées vis-à-vis des indications à propos des niveaux les plus profonds de l'arborescence ou relatives aux éléments individuels (les sous-dossiers et les fichiers) de sorte à pouvoir mener une évaluation *top-down* ou descendante.

Les informations et micro-fonctionnalités jugées pertinentes pour une analyse plus détaillée (coefficient 2) sont les plus nombreuses ; elles sont en outre proposées par un ou deux logiciels différents. Il s'agit alors surtout d'informations relatives aux éléments individuels (comme le poids d'un fichier, ou le nombre de sous-dossiers et fichiers contenus dans un unique dossier) ainsi que des champs dérivés (comme les proportions) et des visualisations graphiques à propos de la volumétrie. La majorité de ces informations et de ces visualisations graphiques peuvent facilement être calculées et générées via des tableurs ou des logiciels de visualisation et ne sont donc pas essentielles à l'évaluation, dans un premier temps du moins. En outre, les informations très détaillées ne sont utiles que si l'on désire mener une évaluation au niveau du fichier, ce qui s'avère le plus souvent impossible au vu du très grand nombre de données transmises aux archives.

Tableau 6 : « Arborescence et volumétrie », par informations et micro-fonctionnalités, résumé

Informations et micro-fonctionnalités	Coefficient	Nombre d'outils offrant cette fonctionnalité	Nom des outils proposant la fonctionnalité
Poids total de répertoire analysé	3	5	Tous les outils
Gestionnaire de fichiers (file manager)	3	4	Tous sauf Archifiltre
Nombre de fichiers d'une extension / d'un groupe	3	4	Tous sauf WinCatalog
Nombre total de fichiers dans le répertoire analysé	3	4	Tous sauf WinCatalog
Poids des fichiers d'une extension / d'un groupe	3	4	Tous sauf WinCatalog
Poids total du répertoire sélectionné (dossiers et fichiers)	3	4	Tous sauf Droid
Statistiques générales sur les formats (répertoire supérieur)	3	4	Tous sauf WinCatalog
Extensions individuelles	3	3	Droid, TSP, WinDirStat
Nombre total de dossiers dans le répertoire analysé	3	3	Archifiltre, Droid, TSP
Nombre total de fichiers subordonnés par branche	3	3	Archifiltre, TSP, WinDirStat
Développement par niveau de l'arborescence	3	2	Droid, TSP
Nombre total de dossiers subordonnés par branche	3	2	TSP, WinDirStat
Statistiques dynamiques sur les formats (par répertoire sélectionné)	3	1	TSP
Statistiques visuelles des formats - surlignage dans l'arborescence	3	1	WinDirStat
Poids d'un fichier	2	5	Tous les outils
Développement complet	2	3	Archifiltre, TSP, WinCatalog
Nombre de fichiers par dossier individuel	2	3	TSP, WinCatalog, WinDirStat
Graphique en barres (Bar Chart)	2	2	TSP, WinDirStat
Groupeement des extensions par types	2	2	Archifiltre, TSP
Proportion vis-à-vis du nombre total de fichiers du dossier parent	2	2	Archifiltre, TSP
Proportion vis-à-vis du poids total du dossier parent	2	2	TSP, WinDirStat
Développement par poids des dossiers	2	1	TSP
Graphique en secteurs (Pie Chart)	2	1	TSP
Graphique en secteurs (Pie chart)	2	1	TSP
Nombre de dossiers subordonnés par dossier individuel	2	1	WinCatalog
Nombre de niveaux de l'arborescence	2	1	Archifiltre
Paramétrabilité des types de fichiers (extensions groupées)	2	1	TSP
Poids moyen des fichiers par branche	2	1	TSP
Stalactites (avec classement et pondération)	2	1	Archifiltre
Stalactites hiérarchiques	2	1	Archifiltre
Statistiques visuelles des formats - surlignage dans la visualisation	2	1	WinDirStat
Carte proportionnelle (Tree Map, par groupe et extension)	1	2	TSP, WinDirStat
Carte proportionnelle volumétrique (Tree Map)	1	2	TSP, WinDirStat
Nombre total d'éléments dans le répertoire analysé	1	2	Droid, WinDirStat
Poids de tous les fichiers dans un dossier individuel	1	2	TSP, WinDirStat
Carte proportionnelle (Tree Map, par fichier et extension)	1	1	WinDirStat
Carte proportionnelle (Tree Map, par répertoire et profondeur)	1	1	TSP
Graphique en barres empilées	1	1	Archifiltre
Nombre total d'éléments par branche (dossiers et fichiers)	1	1	WinDirStat
Nombre total d'éléments par dossier individuel (dossiers et fichiers)	1	1	WinCatalog

### 3.3.2.2 Arborescence et volumétrie : outils

Le test comparatif permet également de sélectionner l'outil (ou les outils) avec le(s)quel(s) une analyse de l'arborescence et de la volumétrie sera la plus pertinente. Nous présentons ci-dessous les caractéristiques des quatre outils les plus intéressants (*WinCatalog* ne sera pas abordé dans le détail en raison du faible score obtenu – 18 –, et du nombre restreint de fonctionnalités offertes – 8 sur 40 ; on invite donc les archivistes intéressés par ce logiciel à consulter le Tableau 8 : Tableau analytique « Arborescence et volumétrie »).

Tableau 7: « Arborescence volumétrie », par outils, résumé

Bilan final	Outils				
	Archifiltre	DROID	TreeSize Professional	WinCatalog	WinDirStat
Informations données	15	11	30	8	22
Score pondéré	37	30	70	18	49
Fonctionnalités pas ou peu utiles (1)	1	1	4	1	6
Fonctionnalités pertinentes (2)	6	1	12	4	5
Fonctionnalités nécessaires (3)	8	9	14	3	11

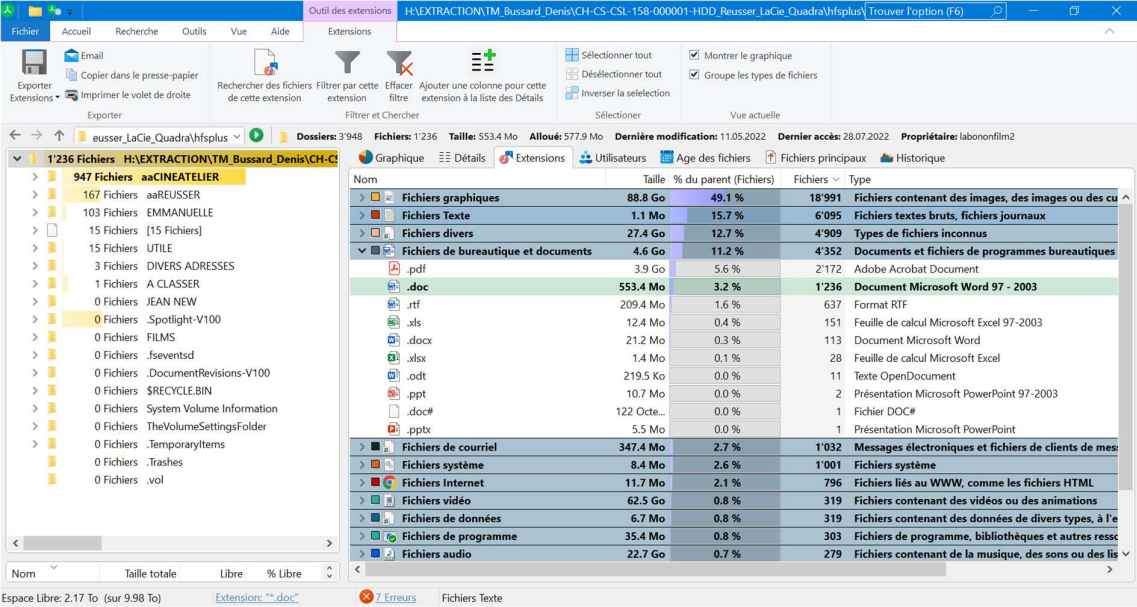
#### 3.3.2.2.1 TreeSize

Comme on peut le voir dans le tableau ci-dessus, c'est le logiciel développé par la société Jam Software depuis 1997 (*TreeSize*) qui paraît le plus indiqué. Ce dernier ne propose pas moins de 30 informations / micro-fonctionnalités sur 40 et obtient un score pondéré de 70. En outre, 26 micro-fonctionnalités proposées sont jugées nécessaires (14 éléments avec le coefficient 3) ou pertinentes (12 avec un coefficient 2) pour mener une évaluation, ce qui le place loin devant les autres logiciels testés.

Parmi les points forts de *TreeSize*, relevons la coexistence d'un gestionnaire de fichiers traditionnel *et* de visualisations graphiques multiples (le gestionnaire est d'ailleurs « disposé » sur un graphique en barres qui apparaît en transparence, et il indique directement le nombre et le poids des éléments contenus aux côtés des noms des répertoires) ; un développement complet *et* sélectif de l'arborescence (permettant notamment de travailler par profondeur et de comparer la structure des différents niveaux) ; des informations quantitatives quasi exhaustives à propos du nombre et du poids des fichiers et des dossiers (pour le répertoire complet, par branche de l'arborescence *et* par dossier individuel). *TreeSize* se distingue aussi tout particulièrement par les informations qu'il fournit sur les formats de fichiers puisqu'il propose des regroupements par types de fichiers (les groupes sont d'ailleurs paramétrables) ainsi qu'une liste d'extensions, avec des indications sur le nombre et le poids des fichiers de chaque extension. Il est enfin le seul logiciel à proposer deux micro-fonctionnalités extrêmement utiles pour la prise de connaissance et le traitement des formats de fichiers : des « Statistiques dynamiques sur les formats (par répertoire sélectionné) » ainsi que des « Statistiques visuelles des formats – surlignage dans l'arborescence ». Il est effectivement possible de générer des statistiques non seulement pour le répertoire entier (de premier niveau, comme presque tous les logiciels le proposent), mais également d'obtenir des informations sur les formats *par répertoire sélectionné*, ce qui permet de connaître la composition exacte de chaque dossier ; en outre, si l'on sélectionne une extension, l'arborescence est modifiée directement et le nombre de fichier de ladite extension s'affiche à côté du nom des répertoires contenant au moins un fichier de ce format. Dans l'image ci-dessous, on voit ainsi apparaître non seulement la liste des formats de fichiers présents dans le Fonds Reusser (par « Type », avec le regroupement de plusieurs formats, ainsi que la liste détaillée des extensions pour chacun des types), mais également une actualisation de

l'arborescence dans le panneau de gauche : le gestionnaire de fichiers indique effectivement aux côtés de chaque nom de dossier le nombre de fichiers d'une extension sélectionnée et classe les dossiers dans l'ordre (l'exemple ci-dessous porte sur la distribution, au sein du Fonds Reusser, des fichiers dont l'extension est « .doc ») :

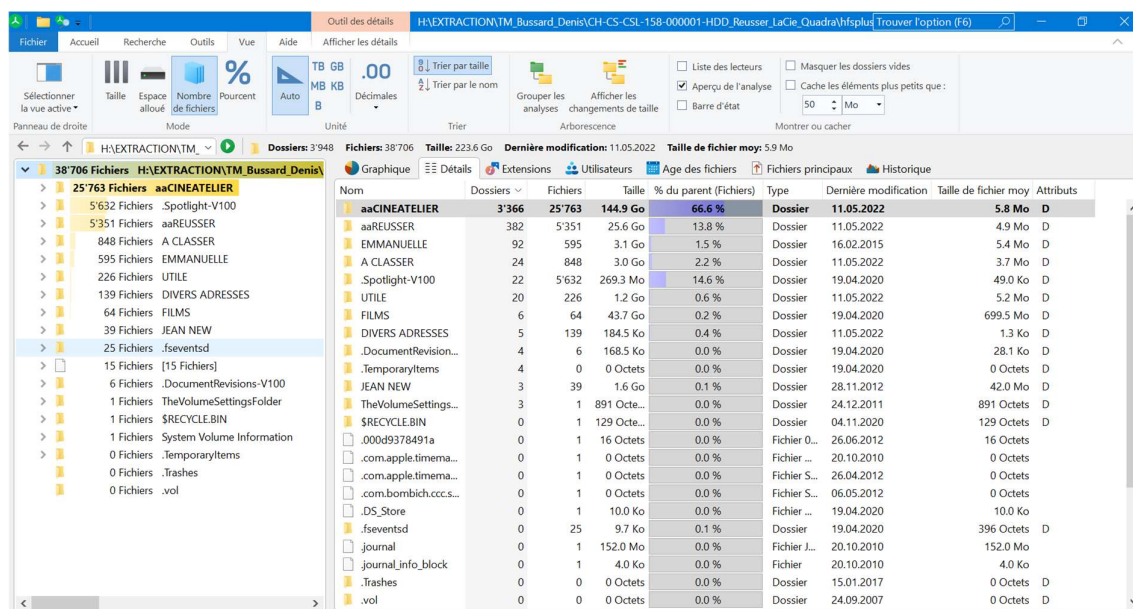
Figure 10 : Vue « Extensions », avec les fichiers « .doc » surlignés dans l'arborescence, *TreeSize*, Fonds Reusser



*TreeSize* se distingue donc surtout par les caractéristiques suivantes : il est hautement paramétrable ; toutes les informations quantitatives sont dynamiques et reflètent donc le contenu du répertoire ou du dossier sélectionné ; et il permet d'obtenir des informations non seulement au niveau des fichiers mais également à propos des dossiers tout en travaillant sur différents niveaux de l'arborescence.

Côté utilisateur, l'interface graphique de *TreeSize* est très conviviale et simple d'utilisation (les champs sont clairement légendés, les extensions sont facilement reconnaissables via des icônes, etc.). Le logiciel présente deux fenêtres principales. Dans l'une figure le gestionnaire de fichiers, dans l'autre se trouvent différentes vues à choix présentant entre autres des graphiques, des informations quantitatives et techniques (sous la dénomination « Détails ») et des statistiques sur les formats de fichiers (onglet intitulé « Extensions »).

Figure 11 : Vue « Détails » avec les éléments classés en fonction du nombre de dossiers, *TreeSize*, Fonds Reusser

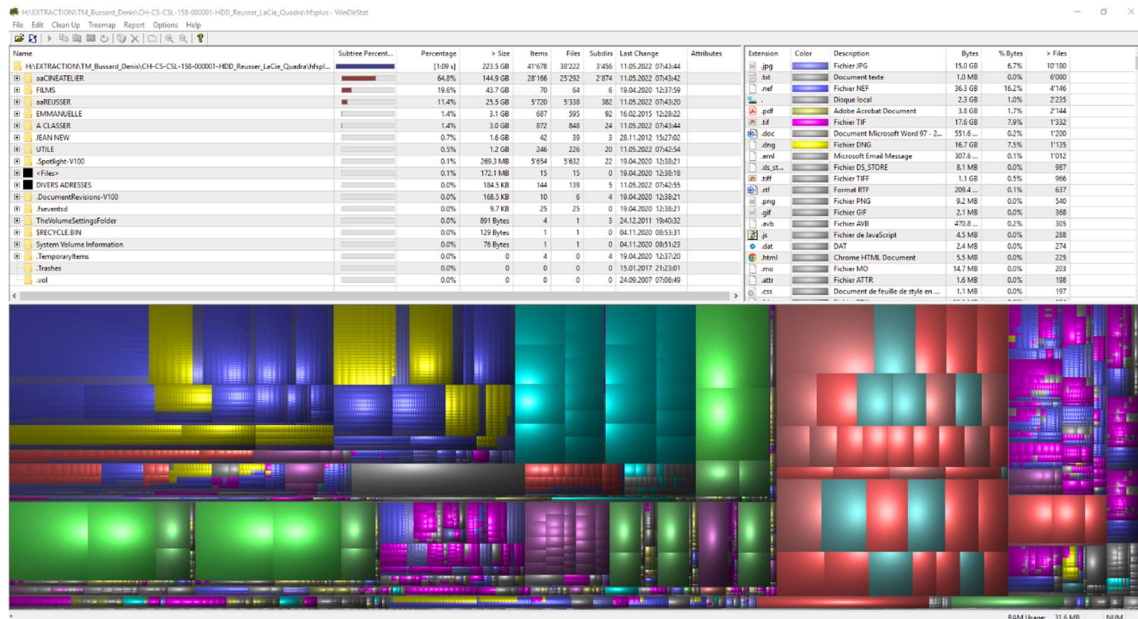


### 3.3.2.2.2 WinDirStat

L'outil *open source* *WinDirStat* occupe le deuxième rang des logiciels testés pour l'analyse de l'arborescence et de la volumétrie avec 22 micro-fonctionnalités et informations proposées, dont 11 sont jugées nécessaires et 5 sont considérées comme pertinentes (et un score pondéré de 49). Le logiciel fonctionne de la manière suivante : il parcourt l'ensemble des données contenues dans le répertoire à analyser et présente les résultats dans trois panneaux distincts. Il y a d'une part un gestionnaire de fichiers livrant des informations dans les colonnes « Nom », « Taille », « Sous-dossiers », « Fichiers », « Éléments », « Pourcentage (graphique) », « Pourcentage », « Dernier changement » et « Attributs ». Si les colonnes ne sont pas paramétrables, le logiciel fournit tout de même de nombreuses informations de nature quantitative sur les fichiers et les dossiers. Il est par exemple le seul avec *TSP* à donner le « Nombre total de dossiers subordonnés par branche » (soit le nombre de sous-dossiers qu'un répertoire contient). Un second panneau contient les données relatives aux extensions de fichiers (description, poids de chaque extension, pourcentage du poids total et nombre de fichiers). Le troisième panneau est propre à ce logiciel puisque lui seul propose une carte proportionnelle (*Tree Map*) représentant la *totalité* des fichiers contenus dans le répertoire principal. Chaque fichier est représenté par un rectangle, dont la couleur correspond à son extension et la surface à son poids relatif.



Figure 12 : Interface unique de WinDirStat, Fonds Reusser



Les trois panneaux communiquent entre eux et sont (partiellement) dynamiques et interactifs, ce qui permet de travailler prioritairement via le gestionnaire, via les extensions, ou via les fichiers sur la carte proportionnelle :

- on peut sélectionner un dossier dans le gestionnaire de fichiers et voir apparaître, dans un rectangle surligné en blanc dans la carte, tous les fichiers dont il est composé ;
- on peut sélectionner une extension dans le panneau de droite et voir apparaître, dans des rectangles surlignés en blanc dans la carte, tous les fichiers de ladite extension à travers *tout* le répertoire analysé ;
- on peut sélectionner un rectangle particulier dans la carte proportionnelle et découvrir aisément à quel groupe d'extension il appartient et dans quel dossier il se trouve dans le gestionnaire.

Cette visualisation « à plat » de l'arborescence (puisqu'elle apparaît ainsi littéralement « dépliée », avec tous les fichiers qu'elle contient sur un même plan) permet de repérer très facilement comment et de quoi le répertoire est composé en termes de formats de fichiers. On identifie aisément les fichiers les plus volumineux (par la taille des rectangles) et les formats les plus représentés (par le nombre de rectangles d'une même couleur). Si ces informations sont également disponibles sous forme statistique (dans le panneau des extensions en haut à droite, de même que dans l'onglet « Extensions » de *TSP* ou dans les rapports de *Droid* par exemple), cette visualisation inédite offre la possibilité de prendre connaissance de l'emplacement et du « voisinage » de chaque fichier (dans quel dossier se trouve un fichier volumineux ? est-il isolé dans l'arborescence, ou appartient-il à un ensemble d'éléments de même poids ? un fichier d'un format particulier côtoie-t-il d'autres fichiers de la même extension, ou est-il distinct de ses voisins directs ?) ; de repérer des agrégats de fichiers de même poids et / ou de même format ; et enfin de mieux connaître la composition interne des dossiers : contiennent-ils des fichiers homogènes (poids et / ou formats identiques) ou



hétérogènes (si les fichiers d'un même dossier sont de natures différentes) ? En permettant de prendre connaissance de la nature des fichiers et de la composition des dossiers, *WinDirStat* offre la possibilité de travailler à plusieurs niveaux. Côté utilisateur, son interface graphique est très intuitive, ce qui le rend facile à prendre en main ; il est en outre totalement gratuit et l'interface est disponible en français.

L'interactivité entre les panneaux est malheureusement limitée. Les informations quantitatives à propos des extensions sont statiques et ne concernent que le répertoire analysé dans son ensemble ; de même, la carte proportionnelle reste inchangée quel que soit le dossier que l'on sélectionne dans le gestionnaire (le dossier est surligné seulement) et affiche donc uniquement la composition du répertoire supérieur (ce qui nécessite de charger un répertoire inférieur et de recommencer le scan complet si on veut une carte proportionnelle plus spécifique). Cette impossibilité de « zoomer » sur certaines parties de l'arborescence rend malheureusement difficile la lecture de la carte proportionnelle pour les supports numériques contenant de très nombreuses données (à quoi s'ajoute un code couleur par extension peu intelligible car il ne correspond à rien de familier pour l'utilisateur des applications les plus communes – bleu pour *Word*, vert pour *Excel*, rouge pour *Acrobat*, etc.). Enfin, au nombre des arguments en défaveur du logiciel, signalons encore que ce dernier est peu paramétrable (les menus sont d'ailleurs très réduits), qu'il n'offre quasiment aucune possibilité de manipuler les données (seule l'élimination est proposée, temporaire ou définitive, et quelques lignes de commande pour l'automatisation de certaines tâches), qu'il ne propose aucun filtre de recherche, ni de rapport synthétique ou d'extraction de métadonnées. *WinDirStat* se révèle donc surtout un bon outil de consultation pour saisir d'un seul coup d'œil la composition d'un répertoire en termes de format et de volume de fichiers.

#### 3.3.2.2.3 *Archifiltre*

Le logiciel *open source* développé par la Fabrique Numérique des Ministères Sociaux de l'État français *par* et *pour* des archivistes obtient un score pondéré de 37, et propose 15 fonctionnalités / informations sur les 40 que compte le tableau analytique (dont 8 sont nécessaires, 6 pertinentes, et 1 jugée pas ou peu utile).

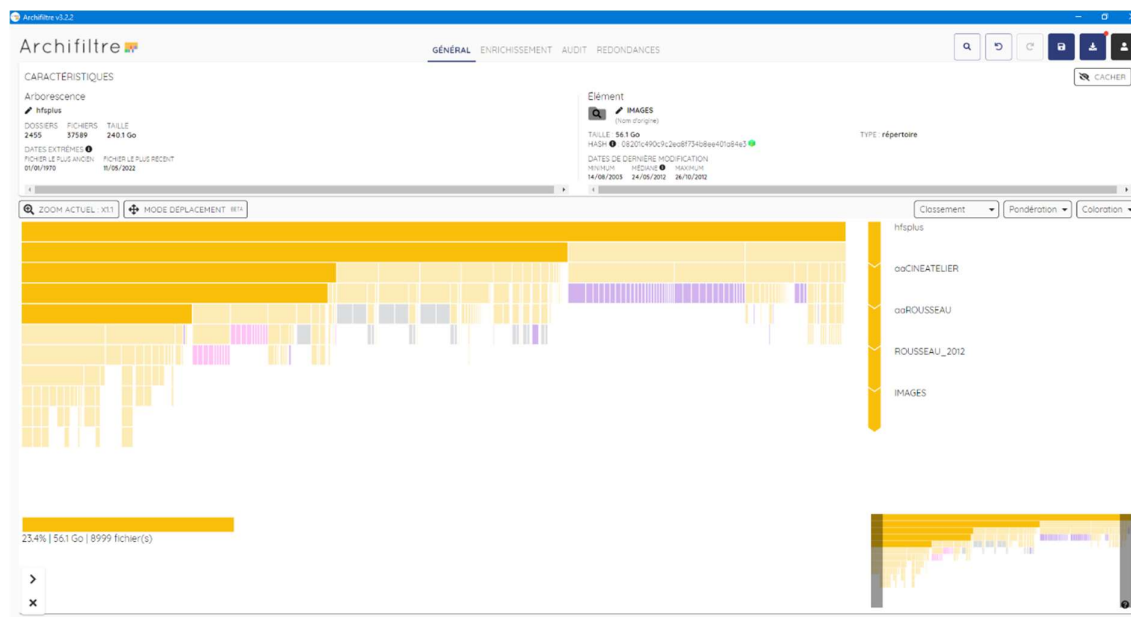
*Archifiltre* se concentre principalement sur l'analyse d'arborescences complètes et offre une visualisation totalement inédite (c'est d'ailleurs le seul outil à ne pas proposer de gestionnaire de fichiers en parallèle) : l'intégralité du répertoire se déploie sous les yeux de l'utilisateur sous la forme de stalactites hiérarchiques, avec les dossiers de premier niveau au sommet (occupant toute la largeur de la fenêtre) et les dossiers qui leur sont directement subordonnés en-dessous. L'arborescence est ainsi « lisible » haut en bas, par niveau de profondeur des dossiers. À cette première lecture, très intuitive, s'ajoutent deux possibilités de filtre : on peut classer les éléments de gauche à droite par volume, par date ou dans l'ordre alphanumérique ; et on peut pondérer la surface des rectangles par volume ou par nombre de fichiers contenus. L'arborescence se lit ainsi de trois façons différentes, selon l'angle d'approche : elle se parcourt de haut en bas pour étudier la hiérarchie des dossiers et leur imbrication ; elle se lit, « par niveau », de gauche à droite, pour découvrir les dossiers les plus lourds / légers, ou les plus anciens / récents ; et elle permet, « par étage », de voir grâce à la surface des rectangles quels sont les dossiers dont le poids ou le nombre de fichiers sont les plus conséquents.

Cette tripartition des vues (verticale, horizontale et par pondération des surfaces) peut parfois brouiller la lecture et il faut toujours garder en tête que la lecture de haut en bas, par

profondeur, est première et prime sur les deux autres : les deux filtres s'appliquent toujours uniquement à l'étage considéré et aux éléments subordonnés à un répertoire (les dossiers tout à gauche de l'écran sont par exemple les plus volumineux à *leur niveau de profondeur* et à *l'intérieur d'un répertoire donné* – le classement ne vaut donc pas pour l'arborescence entière).

Deux onglets concernent directement l'analyse de l'arborescence : l'onglet « Général » (voir Figure 13) composé de trois panneaux (un panneau en haut à gauche avec les informations générales sur le répertoire analysé ; un panneau en haut à droite pour les métadonnées relatives à un élément sélectionné ; et un panneau central contenant la visualisation en stalactites, avec le chemin d'accès de l'élément sélectionné déployé sur le côté droit de la fenêtre) ; et l'onglet « Audit » qui fournit en plus un graphique en barres empilées de la répartition des fichiers par type, et une information quant au nombre de niveaux de l'arborescence (voir Figure 14).

Figure 13 : Onglet « Général » d'*Archifiltre*, classement et pondération par volume, Fonds Reusser

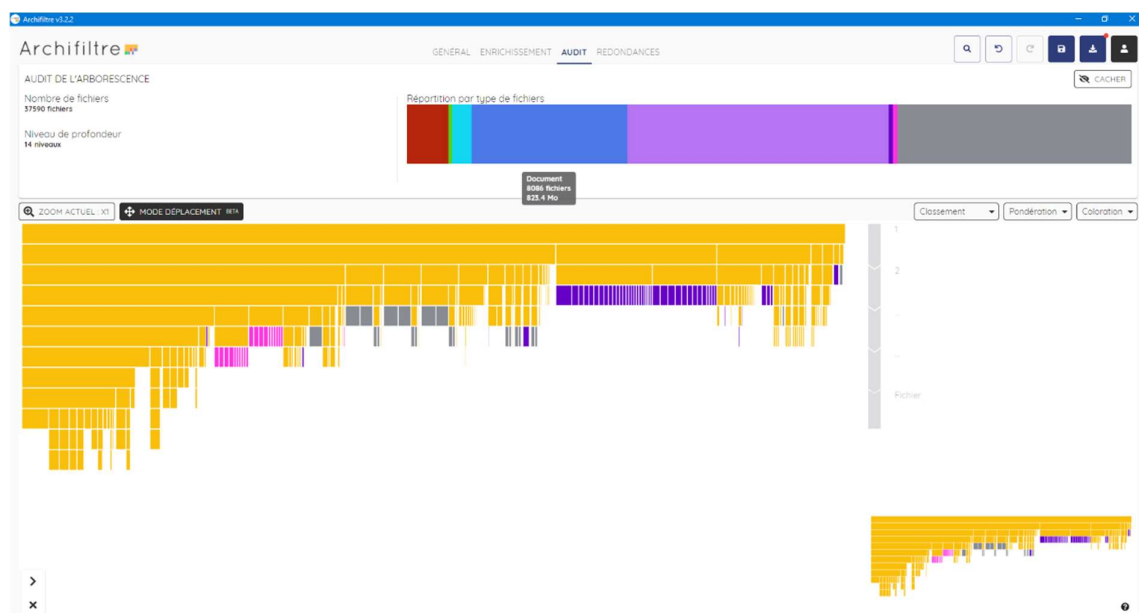


La visualisation en stalactites est très efficace pour prendre connaissance rapidement de la structure générale de l'arborescence : de combien de niveaux de profondeur est fait le répertoire à analyser ? Y a-t-il différents niveaux de profondeur au sein de l'arborescence (à quel endroit, et pour quel type de répertoire) ? Combien de dossiers contient en moyenne un répertoire individuel ? Observe-t-on un classement homogène (avec des structures similaires par répertoire, qui témoigneraient d'une organisation sérielle, chronologique ou thématique par exemple) ou hétérogène (avec des niveaux et un nombre de dossiers très différents d'un répertoire à l'autre) ?

Hormis cet outil de visualisation, *Archifiltre* offre peu d'informations ou de micro-fonctionnalités particulièrement pertinentes ou inédites vis-à-vis de ses concurrents – à l'exception des dates extrêmes d'un dossier et de la date médiane, et du nombre absolu de niveaux de profondeur. En termes statistiques par exemple, le logiciel ne donne que le nombre total de fichiers par branche ainsi que le poids total du répertoire sélectionné (ces informations apparaissent en passant le curseur sur un élément). À cet égard, il peut paraître paradoxal qu'un logiciel qui

travaille principalement sur la base des dossiers (qui donnent réellement forme à l'arborescence et qui apparaissent en jaune dans la visualisation) ne fournisse aucune information quantitative sur le nombre de dossiers contenus dans une branche complète ou dans un répertoire individuel. De même, *Archifiltre*, depuis la version V3.0.0, n'indexe plus les dossiers vides, mais surtout, il ne prend pas en charge les dossiers compressés / archivés : dans le cas du Fonds Reusser, cela représente tout de même 59 dossiers « .zip » qui n'apparaissent pas. Enfin, *Archifiltre* se montre peu prolixe concernant les formats de fichiers : les indications apparaissent uniquement dans l'onglet « Audit », sous la forme d'un graphique en barres empilées (ce dernier n'est d'ailleurs pas légendé – les proportions représentent le nombre de fichiers d'un groupe et non leur poids relatif –, et l'ordre des types de fichiers est arbitraire et immuable, quelle que soit leur répartition à l'intérieur du répertoire analysé). *Archifiltre* travaille d'ailleurs uniquement avec des « types de fichiers » (Publication, Tableur, Email, Document, Image, Vidéo, Audio et Autre dans la vue ci-dessous), et il faut se reporter au rapport d'audit (téléchargeable séparément) pour connaître la composition des groupes (sans que ceux-ci ne soient paramétrables). Enfin, ces informations statistiques sur les types de fichiers (nombre et poids seulement) ne sont ni dynamiques (les informations chiffrées ne sont pas modifiées lorsque l'on sélectionne un répertoire particulier et ne reflètent que le répertoire analysé, contrairement à *TSP*), ni interactives (contrairement à *WinDirStat* qui permet de voir leur répartition dans un dossier particulier).

Figure 14 : Onglet « Audit » d'*Archifiltre* sur les types de fichiers, Fonds Reusser



*Archifiltre* peut donc se révéler utile pour une toute première prise en main d'un vrac numérique afin de déterminer sa structure macroscopique. Travaillant sur les arborescences, il se révèle en revanche peu pertinent pour les fonds qui ne possèdent aucune structure (quel que soit son degré d'organisation), c'est-à-dire pour les collections horizontales de fichiers, dans lesquelles les éléments sont enregistrés côte à côte, sans subdivision hiérarchique (un exemple est donné dans le chapitre 4.2.2.1).

### 3.3.2.2.4 Droid

L'outil développé par le *Digital Preservation department* des Archives nationales du Royaume-Uni (*The National Archives, UK*) propose 11 fonctionnalités (dont 9 jugées nécessaires et 1 pertinente) et obtient un score pondéré de 30, ce qui le place en quatrième position des logiciels les plus utiles pour l'étude de l'arborescence et de la volumétrie.

De prime abord, *Droid* fonctionne en miroir opposé d'*Archifiltre* : on n'y trouve qu'un gestionnaire de fichiers traditionnel avec sept colonnes sur onze relatives à l'identification des formats de fichiers (*Extension, Ids – file format count, Format, Version, Mime type, PUID, Method*) et aucune visualisation graphique de l'arborescence et de la volumétrie. En outre, les informations statistiques générales (nombre et poids du répertoire) ne figurent nulle part dans l'interface principale de l'outil, qui n'affiche que les données relatives aux éléments individuels, en l'occurrence les fichiers, lorsqu'on développe l'arborescence. C'est effectivement une caractéristique de *Droid* : il ne propose aucune donnée à propos des dossiers (seule la date de dernière modification est livrée). Ni le poids total d'un dossier, ni le nombre d'éléments qu'il contient ne sont indiqués, et les sommes de contrôle des dossiers ne sont pas calculées. Cette absence d'informations au niveau des dossiers interdit toute évaluation *top-down* et oblige l'archiviste à travailler au niveau des seuls fichiers.

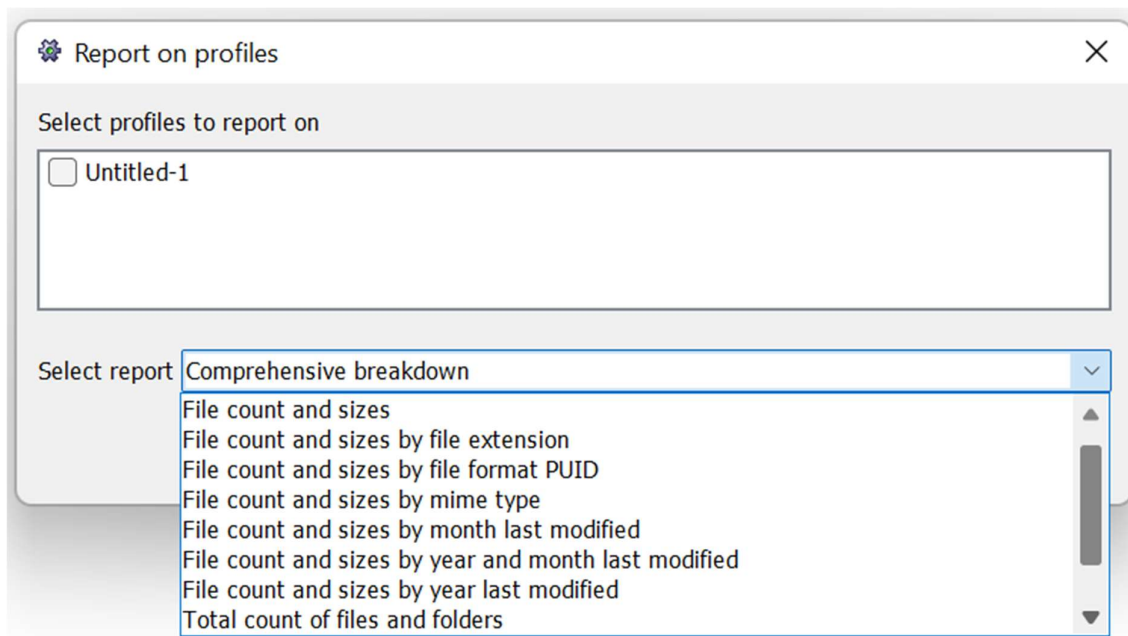
Figure 15 : Interface principale de *Droid*, Fonds Reusser

Extension	Size	Last modified	Id	Format	Version	Mime type	PUID	Method	Hash
PERSONNEL2013		22.03.13 17:39							
PERSONNEL2012		18.09.12 12:36							
PERSONNEL2011		21.01.13 14:54							
PERSONNEL2010_11		16.10.12 15:44							
PERSONNEL2009		07.07.12 11:02							
PERSONNEL2008		06.07.10 13:36							
PERSONNEL2007		03.12.11 11:40							
PERSONNEL2006		13.10.11 11:00							
PERSONNEL2005		19.09.11 21:49							
PERSONNEL2004		23.04.11 17:49							
PERSONNEL2003		22.03.11 11:10							
PERSONNEL2002		16.08.10 17:11							
PERSONNEL2001		19.04.11 09:23							
PERSONNEL2000		23.04.11 17:43							
PERSONNEL1999		12.07.12 19:32							
PERSONNEL1998		06.05.10 11:17							
PERSONNEL1997		13.10.09 15:06							
PERSONNEL1996		19.10.09 11:43							
PERSONNEL1995		11.05.08 13:42							
PERSONNEL1994		13.04.08 12:03							
PERSONNEL1993		19.10.09 11:43							
PERSONNEL1992		23.09.07 15:01							
PERSONNEL1991		09.09.08 19:43							
PERSONNEL1990		31.01.08 14:42							
PERSONNEL1989		13.10.08 19:15							
PERSONNEL1988		25.02.12 19:33							
PERSONNEL1987		07.03.07 18:36							
PERSONNEL1986		17.03.99 10:30							
PERSONNEL1985		11.09.03 10:50							
PERSONNEL1984		07.04.01 14:55							
PERSONNEL1983		28.01.11 11:00							
PERSONNEL1982		19.02.05 11:46							
PERSONNEL1981		16.11.98 16:54							
PERSONNEL1980		26.07.02 10:05							
PERSONNEL1979		24.07.07 17:03							
PERSONNEL1978		20.02.03 17:43							
PERSONNEL1977		22.07.97 11:00							

L'interface principale est d'ailleurs très sobre (voir Figure 15) : les fichiers sont représentés par une icône unique (qui ne reprend donc ni le code couleur de la suite *Microsoft Office* comme dans *Archifiltre*, ni le design propre aux applications à l'instar de ce que propose *TreeSize*) et peu de possibilités de traitement sont offertes (seul un développement sélectif, mais arbitraire, de l'arborescence est proposé : « Expend next three levels »). La meilleure manière d'obtenir des informations plus détaillées passe par l'utilisation des dix rapports synthétiques proposés (voir Figure 16) : il est alors possible de connaître le nombre total d'éléments (dossiers et fichiers) et le nombre total de fichiers (information non disponible pour les dossiers malgré l'utilisation du filtre correspondant) que contient le répertoire analysé. On peut également savoir quel est son poids total, et combien de fichiers de chaque extension, PUID ou Mime-

type il contient. Toutes ces données sont statiques et générales, c'est-à-dire qu'elles se rapportent toujours à l'ensemble du répertoire. Plusieurs rapports ne se révèlent utiles que dans le cadre d'une analyse plus détaillée (« Files count and sizes by year last modified ») tandis que d'autres sont très peu pertinents (« Files count and sizes by month last modified »).

Figure 16 : Rapports proposés par *Droid*



*Droid* propose enfin douze filtres différents, avec les opérateurs de recherche usuels (égal à, plus grand que, etc.). Si les possibilités de recherche n'ont pas été intégrées dans le tableau analytique général ni listées parmi les micro-fonctionnalités proposées, nous accordons une place aux filtres chez *Droid* car l'affichage des résultats présente une particularité qui le rapproche de *TreeSize* (« Statistiques visuelles des formats – surlignage dans l'arborescence ») et qui se révèle intéressante pour l'évaluation : l'arborescence est grisée et seuls les dossiers qui comportent des fichiers répondant aux critères sélectionnés s'affichent. La présence d'éléments correspondant à la recherche est ainsi contextualisée et pas seulement fournie à l'utilisateur sous forme de liste « à plat ».

Figure 17 : Affichage des résultats filtrés, extension « .doc », *Droid*, Fonds Reusser

Extension	Size	Last modifié	Ids	Format	Version	Mime type	PUID	Method	Hash
doc	31 KB	22.01.12 11:57		Microsoft Word Document	97-2003	application/msword	fu1140	Container	8535ca30321425442529...
doc	512 KB	22.06.12 07:52		Microsoft Word Document	97-2003	application/msword	fu1140	Container	698f8d8f91a12b17f629...
doc	185 KB	28.01.10 12:44		Microsoft Word Document	97-2003	application/msword	fu1140	Container	8f6e12775746004e012a...
doc	7 KB	03.02.13 10:47		Microsoft Word Document	6.0/95	application/msword	fu1129	Container	518e799f8ea973b4f77...
doc	46 KB	01.02.10 09:30		Microsoft Word Document	97-2003	application/msword	fu1140	Container	c30ab19123d9b256ca5...
doc	21 KB	31.01.13 10:40		Microsoft Word Document	6.0/95	application/msword	fu1129	Container	178c4638eb0dfebfac1...
doc	8 KB	14.01.13 16:15		Microsoft Word Document	6.0/95	application/msword	fu1129	Container	5f2b88b04645331208b9...
doc	8 KB	14.01.13 16:14		Microsoft Word Document	6.0/95	application/msword	fu1129	Container	9c3d428078a2e0af1327...
doc	11 KB	14.01.13 16:13		Microsoft Word Document	6.0/95	application/msword	fu1129	Container	2eacbee2354876b216...
doc	175 KB	31.08.04 12:55		Microsoft Word Document	97-2003	application/msword	fu1140	Container	1122a22a99b98e4b687...
doc	36 KB	23.06.11 13:28		Microsoft Word Document	97-2003	application/msword	fu1140	Container	945188851355609029...
doc	82 KB	11.12.07 00:33		Microsoft Word Document	97-2003	application/msword	fu1140	Container	1595a2849507a0d04e8...
doc	22 KB	25.12.07 13:19		Microsoft Word Document	97-2003	application/msword	fu1140	Container	b7677881105f101865...
doc	24 KB	21.02.13 12:15		Microsoft Word Document	97-2003	application/msword	fu1140	Container	446321991c1e1f192d...
doc	23 KB	12.02.13 17:37		Microsoft Word Document	97-2003	application/msword	fu1140	Container	c1a3aefc24b27406d...
doc	25 KB	23.12.09 11:08		Microsoft Word Document	97-2003	application/msword	fu1140	Container	5f9e177173a827c787...
doc	34 KB	05.03.07 16:27		Microsoft Word Document	97-2003	application/msword	fu1140	Container	a63136a0e17173b9b...
doc	22 KB	23.12.09 11:13		Microsoft Word Document	97-2003	application/msword	fu1140	Container	c15176a323f1a9b9b...
doc	512 KB	22.06.12 07:52		Microsoft Word Document	97-2003	application/msword	fu1140	Container	698f8d8f91a12b17f629...
doc	185 KB	28.01.10 12:44		Microsoft Word Document	97-2003	application/msword	fu1140	Container	8f6e12775746004e012a...

Comme son nom l'indique (*Digital Record Object Identification*), cet outil est donc spécialisé dans l'identification de formats (il utilise plusieurs méthodes de reconnaissance, par *extension*, *signature* ou *container* et s'appuie sur la base de données PRONOM pour l'analyse de quelque 1'400 formats différents); il travaille donc quasi exclusivement au niveau de l'élément *individuel* (le fichier). En outre, plusieurs manipulations sont nécessaires pour obtenir des informations contextuelles et statistiques générales sur le nombre et le poids de chaque extension. Enfin, le nombre de métadonnées récoltées et affichées dans l'interface est assez restreint (certaines colonnes contiennent en réalité des informations dérivées / secondaires, comme « Method », ou redondantes pour un utilisateur qui ne s'occupe pas spécialement de préservation numérique, comme « Mime type », « PUID », « Version », « Ids »). Pour toutes ces raisons, *Droid* se révèle peu utile pour l'analyse de l'arborescence / volumétrie et ne peut qu'accompagner un autre outil, dans le domaine spécifique de l'identification de fichier non reconnu.

### 3.3.2.3 Arborescence et volumétrie : conclusion

Au terme de cette analyse, force est de constater que les logiciels ont des fonctionnements relativement distincts, mais qu'aucun d'eux ne permet une approche holistique. Ce qui apparaît au fil de cette étude comme une évidence transparaît ici aussi : les logiciels adoptent des points de vue différents sur les répertoires et les arborescences, et traitent en priorité les extensions et les formats (comme *Droid*), s'intéressent aux dossiers qui composent l'arborescence plus qu'aux fichiers individuels (*Archifiltre*), ou privilégient des représentations graphiques et des visualisations originales (comme *WinDirStat*). Seul *TreeSize* semble adopter un point de vue général, en proposant plusieurs approches d'un répertoire : il cumule effectivement l'ergonomie habituelle (un gestionnaire de fichiers), des visualisations graphiques différentes, des pondérations et des tris par nombre ou par poids des éléments selon le point de vue que l'on voudra adopter, des développements sélectifs de l'arborescence par niveau, et des informations détaillées sur les types de fichiers et les extensions. Pour toutes ces raisons, il nous semble pertinent de recommander son acquisition par la Cinémathèque, malgré le fait qu'il s'agisse d'un logiciel propriétaire.

Tableau 8 : Tableau analytique « Arborescence et volumétrie »

Fonctionnalité	Domaines	Informations	Coefficient	Outils										Outils offrant cette fonctionnalité
				Archifiltre		DROID		TreeSize Professional		WinCatalog		WinDirStat		
Arborescence et volumétrie	Visualisation graphique de l'arborescence	Gestionnaire de fichiers ( <i>file manager</i> )	3	0	0	1	3	1	3	1	3	1	3	4
		Stalactites hiérarchiques	2	1	2	0	0	0	0	0	0	0	0	1
		Carte proportionnelle ( <i>Tree Map</i> , par répertoire et profondeur)	1	0	0	0	0	1	1	0	0	0	0	1
		Carte proportionnelle ( <i>Tree Map</i> , par fichier et extension)	1	0	0	0	0	0	0	0	0	1	1	1
	Développement de l'arborescence et informations	Développement par niveau de l'arborescence	3	0	0	1	3	1	3	0	0	0	0	2
		Développement par poids des dossiers	2	0	0	0	0	1	2	0	0	0	0	1
		Développement complet	2	1	2	0	0	1	2	1	2	0	0	3
		Nombre de niveaux de l'arborescence	2	1	2	0	0	0	0	0	0	0	0	1
	Visualisation graphique de la volumétrie (nombre et poids)	Stalactites (avec classement et pondération)	2	1	2	0	0	0	0	0	0	0	0	1
		Graphique en secteurs ( <i>Pie Chart</i> )	2	0	0	0	0	1	2	0	0	0	0	1
		Graphique en barres ( <i>Bar Chart</i> )	2	0	0	0	0	1	2	0	0	1	2	2
		Carte proportionnelle volumétrique ( <i>Tree Map</i> )	1	0	0	0	0	1	1	0	0	1	1	2
	Informations relatives au nombre des dossiers et fichiers	Nombre total de fichiers dans le répertoire analysé	3	1	3	1	3	1	3	0	0	1	3	4
		Nombre total de dossiers dans le répertoire analysé	3	1	3	0	0	1	3	0	0	1	3	3
		Nombre total de fichiers subordonnés par branche	3	1	3	0	0	1	3	0	0	1	3	3
		Nombre total de dossiers subordonnés par branche	3	0	0	0	0	1	3	0	0	1	3	2
		Nombre de fichiers par dossier individuel	2	0	0	0	0	1	2	1	2	1	2	3
		Nombre de dossiers subordonnés par dossier individuel	2	0	0	0	0	0	0	1	2	0	0	1
		Nombre total d'éléments dans le répertoire analysé	1	0	0	1	1	0	0	0	0	1	1	2
		Nombre total d'éléments par branche (dossiers et fichiers)	1	0	0	0	0	0	0	0	0	1	1	1
	Informations relatives au poids des dossiers et fichiers	Nombre total d'éléments par dossier individuel (dossiers et fichiers)	1	0	0	0	0	0	0	1	1	0	0	1
		Poids total de répertoire analysé	3	1	3	1	3	1	3	1	3	1	3	5
		Poids total du répertoire sélectionné (dossiers et fichiers)	3	1	3	0	0	1	3	1	3	1	3	4
		Poids d'un fichier	2	1	2	1	2	1	2	1	2	1	2	5
		Poids moyen des fichiers par branche	2	0	0	0	0	1	2	0	0	0	0	1
		Poids de tous les fichiers dans un dossier individuel	1	0	0	0	0	1	1	0	0	1	1	2
	Traitement des formats de fichiers (types et extensions)	Extensions individuelles	3	0	0	1	3	1	3	0	0	1	3	3
		Groupeement des extensions par types	2	1	2	0	0	1	2	0	0	0	0	2
		Paramétrabilité des types de fichiers (extensions groupées)	2	0	0	0	0	1	2	0	0	0	0	1
	Informations relatives aux formats de fichiers	Poids des fichiers d'une extension / d'un groupe	3	1	3	1	3	1	3	0	0	1	3	4
		Nombre de fichiers d'une extension / d'un groupe	3	1	3	1	3	1	3	0	0	1	3	4
		Proportion vis-à-vis du poids total du dossier parent	2	0	0	0	0	1	2	0	0	1	2	2
		Proportion vis-à-vis du nombre total de fichiers du dossier parent	2	0	0	0	0	1	2	0	0	0	0	1
	Statistiques et visualisation des formats	Statistiques générales sur les formats (répertoire supérieur)	3	1	3	1	3	1	3	0	0	1	3	4
		Statistiques dynamiques sur les formats (par répertoire sélectionné)	3	0	0	0	0	1	3	0	0	0	0	1
		Statistiques visuelles des formats - surlignage dans l'arborescence	3	0	0	1	3	1	3	0	0	0	0	2
Statistiques visuelles des formats - surlignage dans la visualisation		2	0	0	0	0	0	0	0	0	1	2	1	
Graphique en secteurs ( <i>Pie chart</i> )		2	0	0	0	0	1	2	0	0	0	0	1	
Carte proportionnelle ( <i>Tree Map</i> , par groupe et extension)		1	0	0	0	0	1	1	0	0	1	1	2	
Graphique en barres empilées		1	1	1	0	0	0	0	0	0	0	0	1	
Bilan	Nombre d'informations données et score pondéré		15	37	11	30	30	70	8	18	22	49		
	Nombre de fonctionnalités pas ou peu utiles			1		1		4		1		6		
	Nombre de fonctionnalités pertinentes			6		1		12		4		5		
	Nombre de fonctionnalités nécessaires			8		9		14		3		11		



### 3.3.3 Recherche de redondances strictes

Cette section s'intéresse à la manière dont les logiciels traitent les éléments dont il existe des instances strictement identiques à l'intérieur du répertoire analysé. Pour réaliser cette action, les outils informatiques utilisent ce que l'on appelle des *checksums* (on rencontre aussi le terme *hash*). En français, l'on parle alors couramment de « sommes de contrôle », d'« empreintes numériques » ou plus rarement de « condensat ». La méthode consiste en un calcul informatique de cette empreinte qui est le « résultat d'une fonction de hachage appliquée sur une chaîne de caractères de longueur quelconque visant à réduire celle-ci en une donnée de longueur fixe représentative de cette chaîne de caractères. » (Portail International Archivistique Francophone (PIAF) 2015c). Propres à chaque élément puisqu'il s'agit d'[une] « unique alphanumeric value that represents the bitstream of an individual computer file or set of files » (Society of American Archivists 2005), les sommes de contrôle ont avant tout été utilisées dans le domaine de la préservation numérique pour vérifier que les fichiers n'avaient pas été altérés lors de leur transfert (on pouvait effectivement calculer l'empreinte numérique de chaque fichier avant et après un transfert pour voir si la chaîne de caractères qui le compose n'avait pas été modifiée). Mais ces algorithmes de hachage, dont font partie MD5 (*Message Digest 5*), CRC (*Cyclic Redundancy Check*) ou encore SHA (*Secure Hash Algorithm*) et les sommes de contrôle ainsi produites sont également très utiles pour identifier les fichiers qui possèdent « exactement » la même empreinte, c'est-à-dire pour repérer les éléments dont le contenu est le même, quels que soient le nom de l'élément ou sa date de modification par exemple. Nous nous intéressons donc dans cette partie aux logiciels qui utilisent cette fonctionnalité informatique pour identifier les redondances au sein d'un fonds d'archives, et proposer des moyens de les traiter (copie, déplacement, création de raccourcis, élimination, etc.)

#### 3.3.3.1 Recherche de redondances strictes : micro-fonctionnalités et informations

L'analyse des logiciels, via la lecture de leur manuel d'utilisation et par le biais de tests réalisés avec les Fonds Reusser et Simon, nous a permis d'isoler 56 fonctionnalités informatiques dans les domaines généraux suivants : les modes et méthodes de recherche (interface, profondeur, et possibilité de faire des recherches croisées *inter-répertoires* et *intra-dossiers*), le nom de l'algorithme utilisé, l'affichage des résultats (affichage général et détaillé), le type d'informations que le logiciel délivre à propos des redondances (nombre et poids), puis les différents traitements possibles (du mode de sélection des éléments – manuelle ou automatisée – aux actions proposées) et enfin la documentation enregistrée (exportation des résultats et journalisation des événements). Ces domaines et les fonctionnalités subordonnées respectent ainsi l'ordre chronologique d'un traitement de redondances : de la recherche à la documentation de l'action, en passant par l'affichage des résultats et la sélection des éléments.

Dix-huit fonctionnalités (soit un tiers environ) nous paraissent nécessaires (mais pas forcément suffisantes, comme nous le verrons plus tard) : il s'agit entre autres de pouvoir mener une recherche de redondances de fichiers *et* de dossiers à travers tous les répertoires contenus dans un fonds d'archives grâce aux algorithmes qui paraissent les plus fiables et les plus répandus (*MD5* et *SHA-256*). En ce qui concerne l'affichage des résultats, il est hautement préférable que les redondances soient regroupées au sein d'un même onglet et qu'il soit possible de trier les groupes par les différentes colonnes (nom, date, chemin contenant, nombre d'éléments redondants, etc.) afin de pouvoir prioriser le traitement des éléments



multiples. Concernant le traitement proprement dit, le logiciel doit principalement proposer la suppression des éléments (leur déplacement ou leur copie ne semblant guère pertinent dans notre cas) et surtout permettre de sélectionner automatiquement (c'est-à-dire via des filtres prédéfinis ou personnalisés) les éléments que l'on souhaite éliminer. Le nombre de redondances pouvant être très élevé, il n'est effectivement guère envisageable de s'atteler à la sélection manuelle (et donc individuelle) des éléments à traiter (c'est pourtant bien la seule manière de procéder que proposent *Archifiltre* et *WinCatalog*). Dans le domaine archivistique, où l'étude du contexte et le respect des principes de provenance et d'ordre originel des fonds font partie des fondamentaux de la discipline, les logiciels devraient au minimum proposer un traitement des redondances au sein d'un même dossier et plus favorablement encore, au sein d'un même répertoire ou chemin contenant (quels que soient finalement la profondeur et le niveau de précision dudit chemin : d'une branche générale de l'arborescence à un sous-dossier individuel).

**Tableau 9 : « Recherche de redondances strictes », par informations et micro-fonctionnalités, résumé**

Informations et micro-fonctionnalités	Coefficient	Nombre d'outils offrant cette fonctionnalité	Nom des outils proposant la fonctionnalité
Fichiers	3	5	Tous les outils
MD5 Hash	3	5	Tous les outils
Recherche simple dans plusieurs dossiers du même répertoire (chemin)	3	5	Tous les outils
SHA-256 Hash	3	4	Tous sauf Archifiltre
Supprimer les éléments sélectionnés	3	4	Tous sauf Droid
Action directe sur les fichiers originaux	3	3	AIIDup, TSP, WinCatalog
Exportation des résultats de recherche	3	3	AIIDup, TSP, WinCatalog
Nombre d'exemplaires de chaque élément redondant	3	3	AIIDup, Archifiltre, TSP
Onglet de recherche	3	3	AIIDup, TSP, WinCatalog
Regroupement des éléments redondants (éléments dans un « onglet »)	3	3	AIIDup, Archifiltre, TSP
Tri possible des groupes par colonnes	3	3	AIIDup, Archifiltre, TSP
Nombre de groupes d'éléments redondants	3	2	AIIDup, Archifiltre
Par chemin	3	2	AIIDup, TSP
Par dossier spécifique	3	2	AIIDup, TSP
Dossiers	3	1	TSP
Garantie de préservation d'un élément minimum par groupe	3	1	TSP
Informations et paramétrage des colonnes par groupes	3	1	TSP
Sélection automatisée (par filtres personnalisés) des éléments	3	1	TSP
Nombre total d'éléments redondants	2	4	Tous sauf Droid
Poids de chaque élément redondant	2	4	Tous sauf Droid
Journalisation / rapport d'événements	2	3	AIIDup, TSP, WinCatalog
Prévisualisation des éléments	2	3	AIIDup, TSP, WinCatalog
Dédupliquer (création de liens durs, Hard Links)	2	2	AIIDup, TSP
Par date (plus récent ou plus ancien d'un groupe)	2	2	AIIDup, TSP
Poids total des redondances	2	2	AIIDup, Archifiltre
Sélection automatisée (par critères prédéfinis) des éléments	2	2	AIIDup, TSP
SHA-160 Hash	2	2	AIIDup, Droid
Affichage sélectif (tous, aucun ou certains fichiers sélectionnés)	2	1	AIIDup
Byte by byte	2	1	AIIDup
Dédupliquer (création de raccourcis, Shortcuts)	2	1	AIIDup
Nombre de redondances par extension / type	2	1	Archifiltre
Onglet dédié aux redondances	2	1	Archifiltre
Par taille (plus petit, plus grand ou valeur seuil)	2	1	AIIDup
Poids des redondances par extension	2	1	Archifiltre
Recherche croisée dans plusieurs répertoires différents (chemins)	2	1	AIIDup
SHA-384 Hash	2	1	AIIDup
SHA-512 Hash	2	1	AIIDup
Récolement	1	4	Tous sauf AIIDup
Sélection manuelle des éléments	1	4	Tous sauf Droid
Copier les éléments sélectionnés	1	3	AIIDup, TSP, WinCatalog
Déplacer les éléments sélectionnés	1	2	AIIDup, TSP
Inversion de la sélection	1	2	AIIDup, TSP
Liste des éléments redondants (éléments « à plat »)	1	2	Droid, WinCatalog
Par emplacement (premier ou dernier d'un groupe)	1	2	AIIDup, TSP
Poids du groupe d'éléments redondants	1	2	AIIDup, TSP
Proportion du nombre d'éléments redondants par rapport au nombre total	1	2	AIIDup, Archifiltre
Recherche simple dans plusieurs répertoires différents (chemins)	1	2	AIIDup, TSP
Renommer les éléments sélectionnés	1	2	AIIDup, TSP
Supprimer les répertoires vides après l'action	1	2	AIIDup, TSP
CRC32	1	1	WinCatalog
Développement / réduction complet des groupe	1	1	AIIDup
Développement sélectif (tous, aucun ou certains fichiers sélectionnés)	1	1	AIIDup
Par longueur de chemin (plus court ou plus long)	1	1	AIIDup
Par nom (plus court, plus long ou valeur définie)	1	1	AIIDup
Proportion du poids des redondances par rapport au poids total	1	1	Archifiltre
Recherche simple au sein d'un même répertoire et d'un même dossier	1	1	AIIDup

### 3.3.3.2 Recherche de redondances strictes : outils

Comme on peut le voir dans le tableau ci-dessous, deux outils tirent leur épingle du jeu : *AIIDup* et *TreeSize*. Il faut dire qu'ils respectent tous les deux le « cahier des charges » idéal que nous venons de dresser : des onglets de recherche spécifiques, des modes de recherche étendus et multiples, un affichage par onglet, des filtres pour la sélection automatisée, etc. Nous revenons dans la section ci-dessous sur quatre outils en fonction des résultats obtenus et de nos préférences.

Tableau 10 : « Recherche de redondances strictes », par outils, résumé

Bilan final	Outils				
	AllDup	Archifiltre	Droid	TreeSize Professional	WinCatalog
Micro-fonctionnalités proposées	45	18	7	34	17
Score pondéré	87	40	16	75	37
Fonctionnalités pas ou peu utiles (1)	15	4	2	10	5
Fonctionnalités pertinentes (2)	15	6	1	7	4
Fonctionnalités nécessaires (3)	14	8	4	17	8

### 3.3.3.2.1 AllDup

Le logiciel qui sort en tête de notre étude comparative est *AllDup*, un outil gratuit développé par une société basée en Thaïlande, et bien référencé par les différentes listes d'outils existants (*COPTR*, mais aussi le réseau germanophone *nestor* et l'Association des archivistes français le mentionnent). Il propose ainsi 45 fonctionnalités (dont 14 sont jugées nécessaires et 15 sont pertinentes pour le traitement des redondances strictes) et obtient un score pondéré de 87.

*AllDup* se distingue tout particulièrement des autres logiciels actifs dans le même domaine par une méthode de comparaison plus étendue (il ne propose pas moins de cinq algorithmes de hachage de fichiers ainsi qu'une comparaison octet par octet) et un mode de recherche plus fin et précis, qui fait sa force. *AllDup* permet effectivement de sélectionner la manière dont les redondances sont recherchées à travers les répertoires indiqués (soit les chemins contenant généraux au sein desquels opère le logiciel, que *AllDup* appelle « dossiers source ») et leurs sous-dossiers. Quatre modes sont possibles :

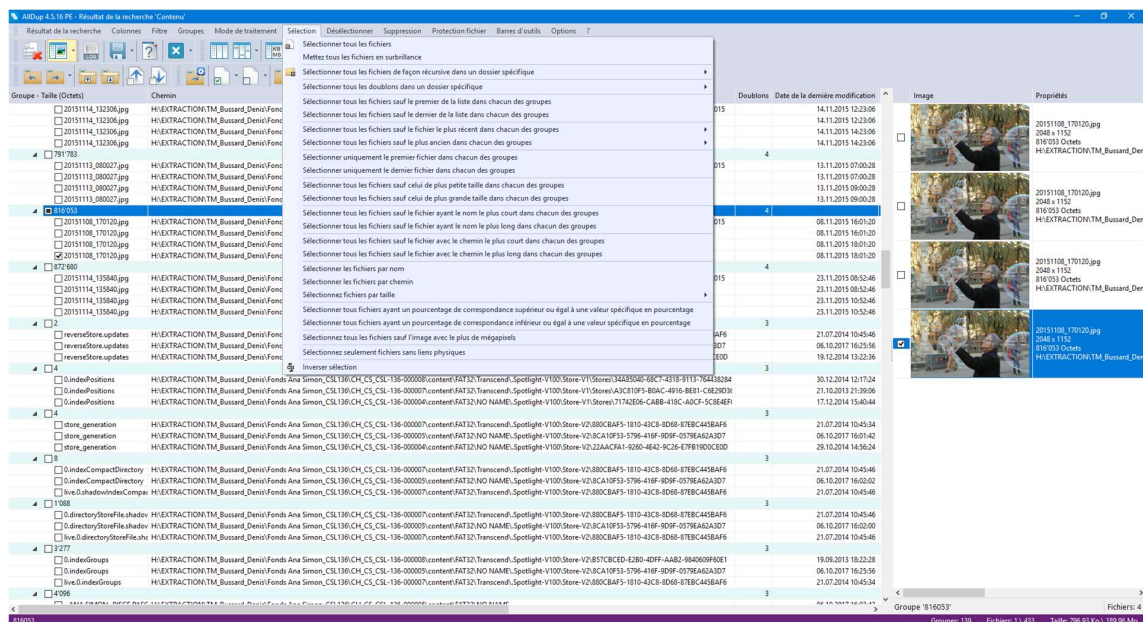
1. Comparer les fichiers de tous les dossiers source
2. Comparer uniquement les fichiers dans le même dossier source
3. Comparer uniquement les fichiers entre les différents dossiers source
4. Comparer uniquement les fichiers d'un même dossier source et d'un même sous-dossier

Si le second mode est commun à tous les logiciels (puisque'il s'agit de ce que nous avons appelé une « Recherche simple dans plusieurs dossiers du même répertoire (chemin) », qui consiste à chercher des redondances dans le répertoire indiqué quel que soit le lieu précis de leur enregistrement, c'est-à-dire dans *tous* les sous-dossiers du répertoire), et que la première option est également proposée par *TreeSize* (soit la possibilité d'ajouter plusieurs chemins contenant / répertoires différents et de lancer une recherche dans *tous* les répertoires indiqués – ce que nous avons appelé « Recherche simple dans plusieurs répertoires différents (chemins) »), *AllDup* est le seul logiciel à notre connaissance qui offre d'une part des recherches « croisées » *inter-répertoires* et d'autre part des recherches limitées aux redondances se trouvant dans le même répertoire *et* le même sous-dossier (*intra-répertoire* et *intra-dossier*). La recherche « croisée » *inter-répertoires* (n° 3) est particulièrement intéressante pour le tri archivistique puisqu'elle permet d'indiquer par exemple deux chemins différents / répertoires et de trouver uniquement les redondances qui se trouvent dans l'un *et* dans l'autre. Cela permet notamment d'identifier les redondances entre des répertoires que nous avons pu repérer précédemment via l'analyse de l'arborescence et de la volumétrie pour leur apparente similarité.

*AllDup* présente aussi l'avantage de proposer de nombreuses façons de filtrer les résultats de recherche et de « traiter » les groupes de redondances : on peut d'une part afficher ou

masquer les groupes, et les développer ou les réduire de manière sélective (en fonction du nombre de fichiers sélectionnés ou l'absence de sélection au sein d'un groupe, ce qui permet de visualiser très rapidement les « groupes » non encore traités), mais aussi sélectionner certains fichiers au sein d'un groupe via des filtres prédéfinis. Si certains sont très peu pertinents (comme la longueur du chemin d'accès ou du nom du fichier, ou encore son emplacement dans la liste – le premier, le dernier, tous sauf le premier, etc.), d'autres s'avèrent particulièrement utiles, voire nécessaires : on peut ainsi sélectionner tous les fichiers qui se trouvent dans un dossier spécifique ou, plus généralement, ceux dont l'enregistrement a été effectué dans un chemin d'accès particulier (permettant ainsi de traiter les éléments non par dossier individuel, mais par branche de répertoire). On peut cependant regretter, chez *AllDup*, le fait que tous les moyens de sélectionner des éléments soient « prédéfinis » comme on peut le voir sur l'image ci-dessous : contrairement à *TreeSize*, il n'existe pas de filtre personnalisé des éléments, soit une option qui permette de sélectionner des éléments en fonction de critères à définir par l'archiviste au cas par cas.

Figure 18 : Recherche de redondances strictes, et filtres de sélection prédéfinis, *AllDup*, Fonds Simon



Signalons encore, à mettre au crédit du logiciel, la possibilité de créer des liens durs (*Hard Links*) ou des raccourcis (*Shortcuts*) pour les éléments redondants (nous reviendrons sur cet aspect dans la présentation de *TreeSize* ainsi que dans la conclusion de ce chapitre sur les redondances) et l'affichage simplifié d'un journal des événements et de statistiques détaillées et de qualité.

### 3.3.3.2.2 *TreeSize*

Chez *TreeSize*, l'identification des redondances se fait dans une fenêtre séparée du gestionnaire de fichiers et des visualisations graphiques (un mode de recherche qui diffère d'*Archifiltre*, mais qui est similaire à celui proposé par *WinCatalog* et qui se rapproche des logiciels spécialisés comme *AllDup* par exemple). Il est effectivement nécessaire d'ouvrir la fenêtre « Recherche de fichier », qui comporte plusieurs modèles prédéfinis en mode avancé (dont « Fichiers volumineux », « Modifié il y a longtemps », « Dossiers vides » ou « Fichiers temporaires »), un onglet « Recherche basique » et l'onglet « Recherche de doublons » (via

notamment le calcul de sommes de contrôle – *TSP* s'appuie sur les algorithmes *MD5* et *SHA256*, soit le plus courant, et l'un des plus précis).

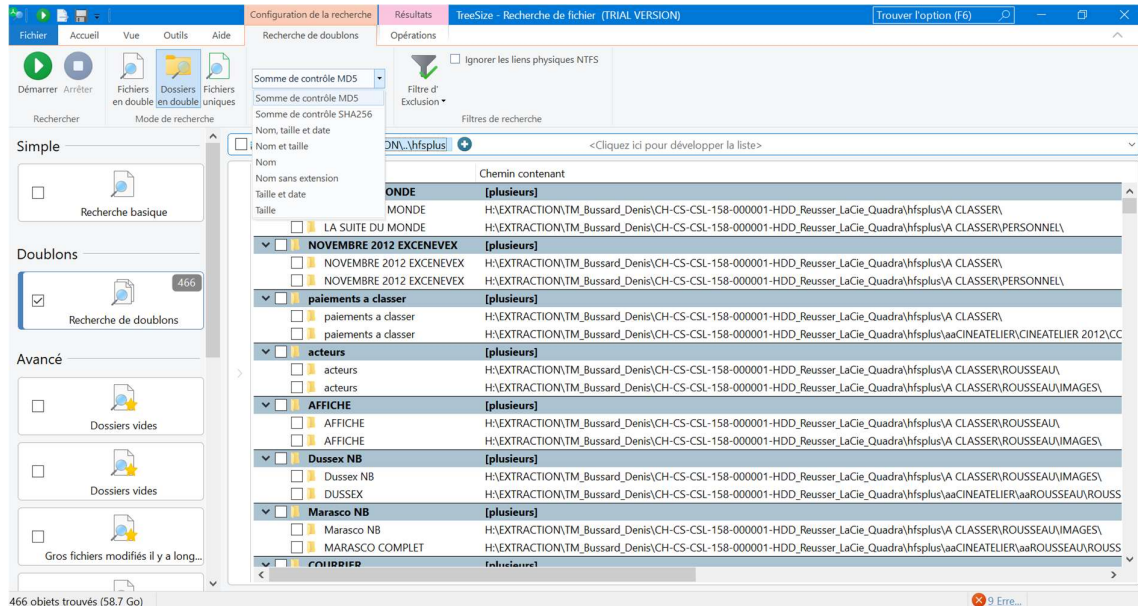
Si *TreeSize* sort deuxième de ce classement, avec 34 micro-fonctionnalités et un score pondéré de 75, il arrive toutefois au premier rang pour ce qui concerne le nombre de micro-fonctionnalités jugées nécessaires (17). L'une des plus importantes, et qu'il est seul à proposer sous cette forme (*Archifiltre* en propose une version incomplète, à notre sens), est la recherche de dossiers redondants (appelée « Dossiers en double » par *TreeSize*<sup>8</sup>). L'identification des dossiers redondants est essentielle pour qui désire mener une évaluation macroscopique : elle permet d'une part de repérer les structures hiérarchiques similaires ; de procéder d'autre part à une évaluation *top-down* (des répertoires, aux dossiers, puis aux sous-dossiers et fichiers) ; et enfin de « dégrossir » le nombre total de fichiers redondants (puisque les dossiers strictement identiques en termes de contenu comportent exactement les mêmes fichiers). Pour toutes ces raisons, la recherche de « Dossiers en double » (via *TreeSize*) est très utile et doit constituer la première étape d'un traitement de redondances au sein d'un répertoire.

*TreeSize* se distingue également par la qualité de l'affichage des résultats. Outre le regroupement des redondances dans un même onglet et l'affichage intégral des résultats sur une seule page (communs à presque tous les outils), *TreeSize* propose des intitulés de colonnes explicites (elles sont également paramétrables) et des titres de groupes immédiatement identifiables (puisque le groupe possède le titre de l'un des éléments subordonnés) ; il permet surtout de trier les ensembles en fonction des colonnes définies, en regroupant les informations multiples sous la dénomination « [plusieurs] ». Si cela peut paraître relativement trivial (bien que cela ne soit pas proposé par *AllDup* ni par *Archifiltre* et encore moins par *WinCatalog*), cette fonctionnalité est d'une grande utilité puisque l'on peut ainsi faire apparaître très facilement les éléments redondants aux caractéristiques semblables ou dissemblables. On peut par exemple travailler prioritairement sur les groupes dont les éléments sont dans le même dossier (ceux dont la colonne « Chemin contenant » indique un chemin d'accès unique et non « [plusieurs] »), ou sur les fichiers redondants encodés dans des formats différents (les groupes dont la colonne « Extension » contient « [plusieurs] »).

---

<sup>8</sup> « Searches for folders that are duplicates of each other. Two folders are considered duplicates, if they contain the same amount of subfolders and files. These subfolders and files also have to be equal to each other, in regards to the selected comparison method » (Jam Software (Joachim Marder) 2022, p. 112)

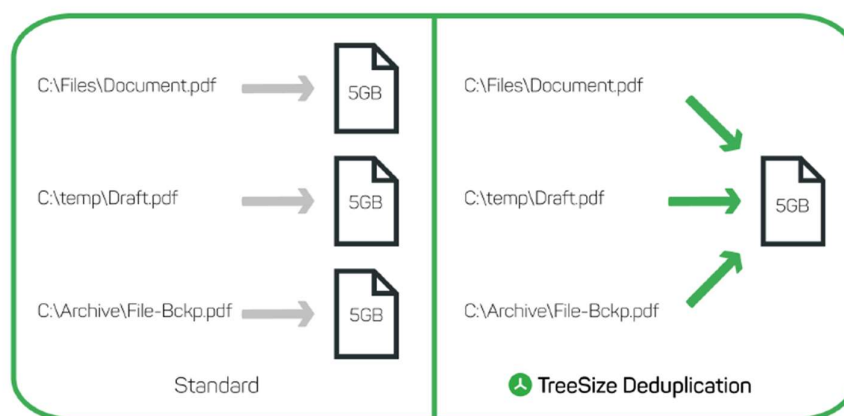
Figure 19 : Recherche de redondances strictes, *TreeSize*, Fonds Reusser



*TreeSize* offre une possibilité de traiter en masse et très rapidement les redondances de *tous* les groupes : il s'agit de la fonctionnalité intitulée sobrement *Deduplication* et qui consiste dans la création de liens durs (*hardlinks*). Le principe expliqué par *TreeSize* dans son manuel d'utilisation est le suivant :

« *The easiest way to gain disk space with the duplicate search is the deduplication feature. Just check the files that you want to deduplicate and select "Deduplicate" from the ribbon menu. TreeSize will replace all but the newest file with NTFS hardlinks. After the deduplication, the copies will no longer allocate space on the drive. [...] Instead of having each of the files take up individual space on your hard disk, TreeSize removes all duplicate files and keeps only one of them. The files that were removed will be replaced by hardlinks, which will then point to the remaining data* » (Jam Software (Joachim Marder) 2022, p. 117)

Figure 20 : Fonctionnement de la « déduplication », par la création de liens durs (*hardlinks*)



JamSoftware

(Jam Software (Joachim Marder) 2022, p. 118)

Comme on peut le voir sur l'image ci-dessus, tous les éléments redondants à l'exception d'un seul (*TreeSize* conserve uniquement le plus récent d'entre eux) sont donc remplacés par un « lien » qui renvoie vers le seul élément conservé. De la sorte, de l'espace disque est économisé, et le contexte (ainsi que le nom) de chaque fichier est préservé. Si cette fonctionnalité est très séduisante (*AllDup* la propose également), nous discutons, dans la conclusion de cette section, de ses implications archivistiques – elle devrait effectivement être utilisée avec parcimonie et de manière sélective.

Entre la sélection manuelle des éléments et la déduplication massive (et aveugle) des éléments redondants via des liens durs, il existe également la possibilité d'automatiser – partiellement du moins – la sélection des éléments : *TreeSize* offre effectivement quelques filtres prédéfinis (à l'instar d'*AllDup*, comme la date des éléments, de sorte à ne conserver que le plus récent), mais surtout il permet de sélectionner les éléments par dossier, par chemin et par des filtres personnalisés. Il est effectivement possible de sélectionner les éléments en fonction des propriétés des fichiers selon des équations de recherche (de type « Et », « Ou », « Égal à », etc.) à définir personnellement et au cas par cas, ce qui fait la grande force de *TreeSize*. L'éventail des possibilités de traitement (de la sélection manuelle au traitement de masse, en passant par la sélection semi-automatisée) est donc large mais *TreeSize* accompagne cependant l'archiviste en lui garantissant qu'*au moins un* élément de chaque groupe de redondances n'a pas été sélectionné (il s'agit d'une option que l'on peut activer dans le menu de *TreeSize* qui permet d'obtenir la garantie de ne pas faire de fausses manipulations qui engendreraient la suppression de *tous* les éléments par mégarde...).

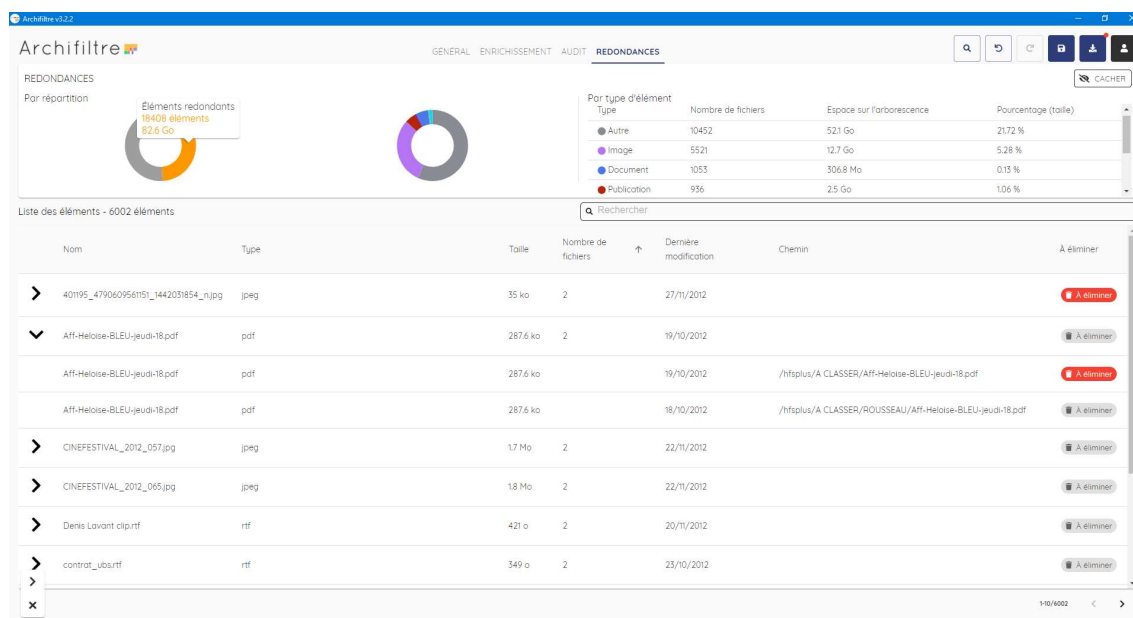
Au nombre des critiques que l'on peut cependant adresser à *TreeSize* figurent notamment les éléments suivants : le logiciel fournit peu d'informations statistiques dans l'interface graphique (seuls le nombre total d'éléments et leur poids sont donnés) et il faut extraire les données pour qu'un résumé quantitatif soit proposé en tête du tableur « .csv » ; *TreeSize* n'offre pas la possibilité de faire des recherches croisées (*inter*-répertoires) ni ciblées (*intra*-dossier) mais seulement d'ajouter plusieurs dossiers et de faire une recherche simple à travers tous les chemins indiqués.

#### 3.3.3.2.3 *Archifiltre*

En ce qui concerne la recherche de redondances strictes, *Archifiltre* se distingue surtout par les statistiques fournies. C'est de loin l'outil qui propose le plus grand nombre de données chiffrées à propos des éléments redondants présents dans le répertoire analysé : le nombre total d'éléments, le nombre de groupes différents, le nombre d'exemplaires de chaque élément de même que le poids total des redondances et le poids de chaque élément sont donnés. En outre, c'est le seul logiciel qui propose une répartition statistique des redondances par type de fichiers (nombre, poids et proportion du poids total). Ces informations quantitatives, qui sont disposées dans les deux panneaux supérieurs de l'onglet, font la particularité d'*Archifiltre* (avec des graphiques en secteurs – malheureusement non légendés –, et un tableau récapitulatif par type) et rendent l'outil intéressant pour une *toute première* prise de connaissance générale des redondances au sein d'un répertoire.



Figure 21 : Onglet « Redondances », *Archifiltre*, Fonds Reusser



En revanche, *Archifiltre* se révèle décevant pour tout ce qui concerne l'identification fine et la navigation dans les résultats de recherche, ainsi que pour le traitement assisté et / ou automatisé des redondances. Parmi les inconvénients du logiciel dans ces trois domaines, citons (dans l'ordre) :

- l'outil fonctionne avec un seul algorithme de hachage (MD5), très commun certes, mais qui n'est pas reconnu comme étant le plus précis ;
- il n'est pas possible d'exporter une liste de tous les éléments redondants dans un tableur (en format « .csv » ou dans *Excel*) – seule une exportation complète du répertoire est possible, à partir de laquelle, moyennant quelques manipulations, on peut accéder à une liste des seules redondances ;
- si le nombre total de dossiers redondants apparaît dans le rapport d'audit téléchargeable séparément, rien ne permet de prendre connaissance des cas concrets et de traiter les dossiers multiples dans l'interface graphique du logiciel ;
- le moteur de recherche spécifique à l'onglet « Redondances » permet de faire des recherches plein texte dans la seule colonne « Nom » (ce qui exclut la recherche par « Chemin » par exemple) ;
- les statistiques et graphiques des panneaux supérieurs ne sont pas dynamiques et ne permettent pas de filtrer les résultats de recherche – le tableau récapitulatif indique d'ailleurs un « type » d'élément par média (image, vidéo, etc.) tandis que la colonne « Type » de la vue détaillée concerne les extensions et les formats individuels (« .csv » par exemple) ;
- le nombre limité de résultats (10) par page ne facilite pas la navigation et empêche toute vue d'ensemble, surtout pour les répertoires volumineux contenant beaucoup de redondances (la Figure 21 montre ainsi les 10 premiers groupes de redondances sur... 6002 que compte le Fonds Reusser) ;
- il n'existe aucune possibilité de visualiser les éléments redondants (que ce soit via un volet de prévisualisation, ou en permettant l'ouverture facilitée d'un explorateur de fichiers) ;



- le tri par colonne (sur les groupes et non sur les éléments) se révèle quasiment inopérant puisque le logiciel ne fournit aucune information sur les chemins au niveau du groupe par exemple (ce qui ne permet pas de voir apparaître les redondances qui se trouvent dans le même répertoire) ou sélectionne une seule valeur lorsque plusieurs éléments ont des caractéristiques différentes (au niveau du groupe, *Archifiltre* indique que la date de modification la plus récente, ou une seule extension pour deux fichiers de formats distincts, etc.) ;
- la sélection des éléments à traiter (pour la suppression) est entièrement manuelle (la seule opération « de masse » proposée par *Archifiltre* consiste à sélectionner *toutes* les occurrences d'un groupe de redondances<sup>9</sup>).

#### 3.3.3.2.4 *Droid*

Il peut paraître surprenant de voir figurer l'outil *Droid* (dédié à l'identification de format dans une démarche de préservation numérique) dans le tableau analytique consacré au dédoublonnage, puisque son interface graphique ne propose aucune fonctionnalité *directe* pour procéder à l'identification, au traitement et à la suppression des éléments redondants – on notera d'ailleurs que si la colonne « Hash » figure dans l'interface, il s'agit de l'une des seules colonnes sur lesquelles on ne peut pas appliquer de filtre. Si nous avons fait le choix de le présenter dans cette section relative à la recherche de redondances, c'est que l'aide en ligne du logiciel (de même que le manuel utilisateur) comporte une rubrique consacrée aux fichiers redondants, « Detecting duplicate files », qui expose les méthodes disponibles : chercher les dossiers dont le nom comporte par exemple « backup », « temp » ou « old » ; examiner les noms de fichiers (s'ils sont similaires ou identiques) pour comparer les contenus ; et générer des sommes de contrôle pour faire apparaître les valeurs redondantes. Cette dernière méthode est la plus fiable et la moins chronophage, et *Droid* la propose en priorité. Pour le traitement proprement dit, la méthodologie indiquée est très rudimentaire et « manuelle » puisqu'il s'agit d'exporter l'ensemble des métadonnées dans un tableur et d'utiliser la fonction de mise en forme conditionnelle (dans *Excel* par exemple) pour surligner et faire apparaître visuellement les fichiers ayant la même valeur de hachage. *Droid* ne fait d'ailleurs pas mystère des limites de sa méthodologie : « DROID can generate content hashes for your files, but note that DROID will not locate files with the same hash value for you, only generate them in the first place » (The National Archives [sans date]). Il nous semblait donc intéressant de présenter également un outil qui se situe à l'autre extrémité sur l'échelle de l'ergonomie et de l'« expérience utilisateur » et proposant une approche totalement différente : ni interface, ni sélection automatisée.

#### 3.3.3.3 Recherche de redondances strictes : conclusion

Au terme des analyses de fonctionnalités et de logiciels, nous pouvons essayer de dresser le portrait-robot de l'outil idéal. Ce dernier devrait proposer : les modes de recherche (croisée ou simple, dans un ou plusieurs répertoires), le journal des événements et les prévisualisations de fichiers d'*AllDup* ; les statistiques d'*Archifiltre* ; l'affichage et l'exportation des résultats ainsi que la sélection assistée des éléments à traiter (en particulier par chemin et par dossier spécifique) de *TreeSize*.

---

<sup>9</sup> On lit ainsi dans le manuel utilisateur : « L'élément redondant peut être tagué « à éliminer », si l'ensemble des redondances sont à éliminer, il est possible d'appliquer le tag sur la ligne supérieure pour l'appliquer aux lignes inférieures » (Fabrique numérique des Ministères Sociaux 2021).

Le portrait-robot de l'outil idéal ne fait pas mention du remplacement des fichiers et dossiers redondants par des liens durs (*Hardlinks*) ou des raccourcis (*Shortcuts*), comme le proposent *AllDup* et *TreeSize*. Si cette fonctionnalité peut certes être utile, plusieurs réserves s'imposent, d'ordre technique, mais aussi archivistique. D'un point de vue technique tout d'abord, les liens durs ne sont pas toujours reconnus par les systèmes de fichiers traditionnels, comme Belovari en a malheureusement fait l'expérience en 2017 :

*« However, Windows Explorer and other systems do not recognize hard links. For them, hard links look and behave like the original duplicate files; they cannot detect their presence or number and thus continue to show original disk usage. Folders with hard links, therefore, still appear to include all files including the original duplicates, and the archivists would qualitatively appraise and process the original number of files, be they duplicates or not. » (Belovari 2017, p. 71)*

Nous avons été confronté à la même difficulté en testant les liens durs de *TreeSize* sur un échantillon : aucune économie d'espace disque n'a été constatée suite à la déduplication en masse via des liens durs proposée par *TreeSize*. En outre, des considérations archivistiques restreignent l'utilisation généralisée de cette technologie. En procédant de la sorte, on peut certes économiser des capacités de stockage, mais on ne propose aucune solution à la problématique du repérage des documents : la recherche par un utilisateur générera un nombre toujours aussi conséquent de résultats puisque *tous* les éléments dédupliqués conservent leurs propriétés et seront affichés (or leur contexte individuel n'est pas toujours pertinent et leur affichage parmi les résultats de recherche produira énormément de « bruit » en sortie et un taux de précision au repérage plus faible). Au niveau intellectuel, traiter en masse les éléments redondants par la création de liens durs revient finalement à « tout garder » et à ne pas assumer l'une des tâches essentielles de l'archiviste : porter un jugement sur la pertinence, l'adéquation à la politique de collection et la valeur patrimoniale des documents créés par un organisme public ou proposé par une personne ou une organisation privée. Enfin, l'utilisation généralisée de liens durs n'est qu'une manière peu habile de repousser dans le temps le traitement concret des redondances au sein de l'arborescence puisque la problématique réapparaîtra avec force lors des prochaines étapes ou fonctions archivistiques : lors d'une évaluation fine des contenus, lors de l'élaboration d'un plan de classement, ou lors de la description des éléments. Par conséquent, les liens durs ou les raccourcis ne devraient être utilisés qu'avec parcimonie, après l'analyse de la logique interne du répertoire analysé (où se trouvent les redondances ? quels liens les redondances entretiennent-elles ? à quel niveau se situent-elles ? etc.) et pour les seuls éléments dont nous souhaitons conserver deux ou plusieurs occurrences car leur contexte l'impose.

Hormis quelques différences dans le nombre de redondances trouvées (voir « Tableau 21 : Recherche de redondances strictes : cas pratiques, résumé »), on peut remarquer que la majorité des logiciels fonctionnent de la même manière : la recherche de redondances concerne principalement les fichiers (plus rarement les dossiers), elle s'effectue grâce aux sommes de contrôle, l'affichage se fait par items regroupés dans un onglet, la sélection est manuelle ou semi-automatisée (par filtres définis ou personnalisés) et le traitement consiste principalement dans l'élimination des redondances (voire leur remplacement par des liens durs ou des raccourcis). Tous les outils se rejoignent donc sur les tâches générales que sont : l'identification, l'affichage des résultats, la sélection et le traitement. En revanche, ce qui fait défaut, et cela de manière générale, ce sont les capacités et les instruments d'analyse des résultats (pas uniquement quantitative) permettant de prendre des décisions éclairées qui respectent la déontologie archivistique (on pense en particulier aux principes de provenance

et de respect de l'ordre originel, ainsi que l'importance accordée au contexte dans lequel s'inscrivent les documents). Nous proposerons donc, dans la section « 4.4 Traitement des redondances strictes », quelques instruments et méthodes pour l'analyse des redondances, à partir des exemples fournis par les fonds Francis Reusser et Ana Simon.

Tableau 11 : Tableau analytique « Recherche de redondances strictes »

Fonctionnalité	Domaines	Micro-fonctionnalités	Coefficient	Outils										Outils offrant cette fonctionnalité
				AllDup		Archifiltre		DROID		TreeSize Professional		WinCatalog		
				Codage	Score	Codage	Score	Codage	Score	Codage	Score	Codage	Score	
Recherche de redondances strictes	Interface de recherche des redondances	Onglet de recherche	3	1	3	0	0	0	0	1	3	1	3	3
		Onglet dédié aux redondances	2	0	0	1	2	0	0	0	0	0	0	1
		Récolement	1	0	0	1	1	1	1	1	1	1	1	4
	Profondeur de la recherche de redondance	Dossiers	3	0	0	0	0	0	0	1	3	0	0	1
		Fichiers	3	1	3	1	3	1	3	1	3	1	3	5
	Mode de recherche (intra / extra répertoires et dossiers)	Recherche simple dans plusieurs dossiers du même répertoire (chemin)	3	1	3	1	3	1	3	1	3	1	3	5
		Recherche croisée dans plusieurs répertoires différents (chemins)	2	1	2	0	0	0	0	0	0	0	0	1
		Recherche simple dans plusieurs répertoires différents (chemins)	1	1	1	0	0	0	0	1	1	0	0	2
		Recherche simple au sein d'un même répertoire et d'un même dossier	1	1	1	0	0	0	0	0	0	0	0	1
	Calcul d'empreintes numériques (hash, checksum)	MD5 Hash	3	1	3	1	3	1	3	1	3	1	3	5
		SHA-256 Hash	3	1	3	0	0	1	3	1	3	1	3	4
		SHA-160 Hash	2	1	2	0	0	1	2	0	0	0	0	2
		SHA-384 Hash	2	1	2	0	0	0	0	0	0	0	0	1
		SHA-512 Hash	2	1	2	0	0	0	0	0	0	0	0	1
		Byte by byte	2	1	2	0	0	0	0	0	0	0	0	1
		CRC32	1	0	0	0	0	0	0	0	0	1	1	1
	Affichage général des résultats	Regroupement des éléments redondants (éléments dans un « onglet »)	3	1	3	1	3	0	0	1	3	0	0	3
		Liste des éléments redondants (éléments « à plat »)	1	0	0	0	0	1	1	0	0	1	1	2
	Affichage détaillé des résultats	Tri possible des groupes par colonnes	3	1	3	1	3	0	0	1	3	0	0	3
		Informations et paramétrage des colonnes par groupes	3	0	0	0	0	0	0	1	3	0	0	1
		Affichage sélectif (tous, aucun ou certains fichiers sélectionnés)	2	1	2	0	0	0	0	0	0	0	0	1
		Prévisualisation des éléments	2	1	2	0	0	0	0	1	2	1	2	3
		Développement / réduction complet des groupe	1	1	1	0	0	0	0	0	0	0	0	1
		Développement sélectif (tous, aucun ou certains fichiers sélectionnés)	1	1	1	0	0	0	0	0	0	0	0	1
		Nombre de groupes d'éléments redondants	3	1	3	1	3	0	0	0	0	0	0	2
	Informations relatives au nombre d'éléments redondants	Nombre d'exemplaires de chaque élément redondant	3	1	3	1	3	0	0	1	3	0	0	3
		Nombre total d'éléments redondants	2	1	2	1	2	0	0	1	2	1	2	4
		Nombre de redondances par extension / type	2	0	0	1	2	0	0	0	0	0	0	1
	Informations relatives au poids des éléments redondants	Proportion du nombre d'éléments redondants par rapport au nombre total	1	1	1	1	1	0	0	0	0	0	0	2
		Poids total des redondances	2	1	2	1	2	0	0	0	0	0	0	2
		Poids de chaque élément redondant	2	0	0	1	2	0	0	0	0	0	0	1
		Poids des redondances par extension	2	0	0	0	2	0	0	0	0	0	0	1
		Poids du groupe d'éléments redondants	1	1	1	0	0	0	0	1	1	0	0	2
		Proportion du poids des redondances par extension par rapport au poids total	1	0	0	1	1	0	0	0	0	0	0	1
	Mode de sélection des éléments	Sélection automatisée (par filtres personnalisés) des éléments	3	0	0	0	0	0	0	1	3	0	0	1
		Garantie de préservation d'un élément minimum par groupe	3	0	0	0	0	0	0	1	3	0	0	1
		Sélection automatisée (par critères prédéfinis) des éléments	2	1	2	0	0	0	0	1	2	0	0	2
	Sélection automatisée (par critères prédéfinis, tous sauf ou uniquement)	Sélection manuelle des éléments	1	1	1	1	1	0	0	1	1	1	1	4
		Par chemin	3	1	3	0	0	0	0	1	3	0	0	2
		Par dossier spécifique	3	1	3	0	0	0	0	1	3	0	0	2
		Par taille (plus petit, plus grand ou valeur seuil)	2	1	2	0	0	0	0	0	0	0	0	1
		Par date (plus récent ou plus ancien d'un groupe)	2	1	2	0	0	0	0	1	2	0	0	2
		Par longueur de chemin (plus court ou plus long)	1	1	1	0	0	0	0	0	0	0	0	1
		Par nom (plus court, plus long ou valeur définie)	1	1	1	0	0	0	0	0	0	0	0	1
		Par emplacement (premier ou dernier d'un groupe)	1	1	1	0	0	0	0	1	1	0	0	2
		Inversion de la sélection	1	1	1	0	0	0	0	1	1	0	0	2
	Actions	Action directe sur les fichiers originaux	3	1	3	0	0	0	0	1	3	1	3	3
		Supprimer les éléments sélectionnés	3	1	3	1	3	0	0	1	3	1	3	4
		Dédupliquer (création de liens durs, <i>Hard Links</i> )	2	1	2	0	0	0	0	1	2	0	0	2
		Dédupliquer (création de raccourcis, <i>Shortcuts</i> )	2	1	2	0	0	0	0	0	0	0	0	1
		Copier les éléments sélectionnés	1	1	1	0	0	0	0	1	1	1	1	3
		Déplacer les éléments sélectionnés	1	1	1	0	0	0	0	1	1	0	0	2
		Renommer les éléments sélectionnés	1	1	1	0	0	0	0	1	1	0	0	2
	Documentation	Supprimer les répertoires vides après l'action	1	1	1	0	0	0	0	1	1	0	0	2
		Exportation des résultats de recherche	3	1	3	0	0	0	0	1	3	1	3	3
		Journalisation / rapport d'événements	2	1	2	0	0	0	0	1	2	1	2	3
	Bilan final	Nombre absolu de micro-fonctionnalités et score pondéré		45	87	18	40	7	16	34	75	17	37	
		Nombre de fonctionnalités pas ou peu utiles			15		4		2		10		5	
		Nombre de fonctionnalités pertinentes			15		6		1		7		4	
		Nombre de fonctionnalités nécessaires			14		8		4		17		8	

### 3.3.4 Comparaison de données

#### 3.3.4.1 Comparaison de données : familles et caractéristiques

La comparaison de données est un (très) vaste sujet, qui mériterait sans aucun doute une étude appropriée de grande ampleur. Dans le cadre de ce Mémoire de Master, nous avons décidé d'approcher trois types de logiciels qui proposent des comparaisons de données via des méthodes très différentes, raison pour laquelle nous n'avons pas établi de tableau analytique comparatif, comme nous l'avons fait pour les trois autres fonctionnalités. Les familles se distinguent principalement par trois caractéristiques.

Leur **mode de recherche** tout d'abord : alors que certains logiciels permettent d'effectuer une recherche *générale* d'éléments à l'intérieur d'un répertoire via des filtres prédéfinis – le contenu des fichiers via des sommes de contrôle et des algorithmes de comparaison ou les métadonnées associées aux éléments –, d'autres outils fonctionnent selon une démarche volontariste et bien plus consciente : on doit préciser au logiciel quels sont les deux ou trois éléments (répertoires, dossiers ou fichiers) que l'on souhaite comparer. *AllDup*, *AntiDupl*, *TreeSize*, et *WinCatalog* fonctionnent selon le premier modèle (on peut ainsi demander à ces logiciels de comparer *indistinctement* les uns avec les autres tous les éléments qui se trouvent dans la branche de l'arborescence indiquée, quel que soit leur nombre). À l'inverse, *Beyond Compare* et *WinMerge* fonctionnent selon le second modèle : il est nécessaire d'indiquer à ces deux outils quels sont les deux (*Beyond Compare*) ou trois (*WinMerge*) éléments (répertoires entiers, dossiers, ou fichiers) que l'on souhaite comparer. Ces modes de recherche influencent donc le type de résultats présentés et la manière dont ils sont affichés.

L'**affichage des résultats de recherche** diffère en fonction de leur mode de recherche : alors que les logiciels qui proposent une recherche générale (*AllDup*, *AntiDupl*, *TreeSize* et *WinCatalog*) n'affichent que les résultats qui correspondent aux filtres, c'est-à-dire, généralement, les fichiers qui sont identiques ou similaires (seul *TreeSize* propose une liste de « Fichiers uniques » dans l'onglet réservé à la « Recherche de doublons »), *Beyond Compare* et *WinMerge* affichent l'entier des résultats, que les éléments soient identiques ou non. Pour les répertoires, on connaît ainsi quels sont les dossiers strictement redondants, mais aussi quels sont les dossiers qui contiennent des éléments uniques ou plus récents vis-à-vis des dossiers avec lesquels ils sont comparés ; pour les fichiers, on peut comparer deux éléments, et voir apparaître leurs ressemblances et leurs dissemblances. En d'autres termes, si la première famille de logiciels n'affiche que les résultats qui coïncident avec le filtre de recherche, la seconde famille livre les résultats *entiers* de la comparaison, qu'il y ait coïncidence ou non (libre ensuite à l'utilisateur de parcourir les résultats et de les interpréter).

Enfin, la **profondeur** de la comparaison peut varier : on peut comparer les métadonnées des éléments (comme *TreeSize* ou *WinCatalog* le font par exemple via la recherche des noms, des tailles, ou des dates identiques de dossiers et fichiers) ou bien l'on peut comparer les contenus des fichiers via des algorithmes (*Mean Square Difference* ou *Index of Structural Similarity*, *SSIM*, comme le fait *AntiDupl*, ou via les algorithmes de hachage *aHash*, *bHash*, *dHash* et *pHash* proposés par *AllDup*), des sommes de contrôle ou des comparaisons binaires (comme le font par exemple *Beyond Compare* et *WinMerge* pour les comparaisons de dossiers et de leurs contenus). Il faut préciser ici qu'un logiciel peut utiliser plusieurs méthodes : *TreeSize* et *WinCatalog* utilisaient bel et bien des sommes de contrôle pour la reconnaissance d'éléments *strictement* identiques, mais ces deux outils nous intéressent dans cette section consacrée aux comparaisons de données pour leur faculté à comparer et afficher les éléments

qui présentent des métadonnées identiques ; de même, *Beyond Compare* et *WinMerge* utilisent des algorithmes pour la comparaison fine de deux images et la comparaison par blocs pour les fichiers texte, mais ce sont leurs capacités à comparer le contenu de deux ou trois répertoires et dossiers et d'afficher l'entier des résultats – les éléments identiques comme les éléments divergents – qui nous intéressent maintenant).

Il y aura donc une gradation dans le niveau de comparaison : si *Beyond Compare* et *WinMerge* sont finalement un prolongement du traitement des redondances strictes (puisqu'ils ne permettent pas d'afficher les éléments *similaires*, mais seulement les éléments identiques ou divergents) par la comparaison de répertoires entiers, *TreeSize* et *WinCatalog* offrent une première approche dans l'étude des fichiers similaires (par leurs métadonnées seulement) tandis que *AllDup* et *AntiDup* représentent sûrement la dernière étape (la plus détaillée, via des algorithmes qui analysent le contenu d'un type spécifique de fichiers, en l'occurrence les images).

### 3.3.4.2 Comparaison de données : outils

En nous basant sur les différents types d'outils présentés ci-dessus (par mode de recherche, par affichage des résultats et par profondeur de la comparaison), et sur la gradation dans le niveau de comparaison (des répertoires identiques ou différents aux fichiers images en passant par les seules métadonnées), nous avons sélectionné trois outils qui seront présentés dans les chapitres à venir : *Beyond Compare*, *TreeSize* et *AntiDupl*.

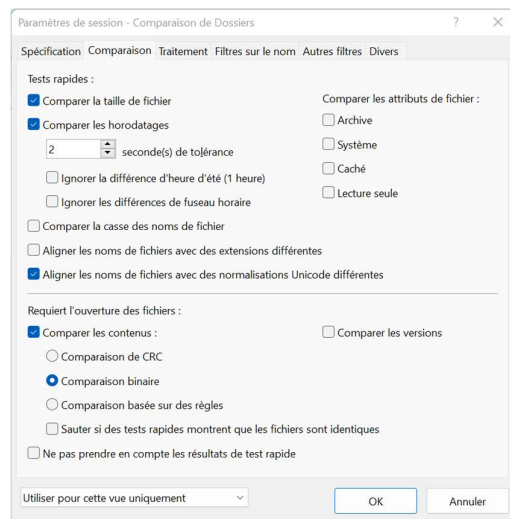
#### 3.3.4.2.1 *Beyond Compare*

Comme son nom l'indique, le logiciel propriétaire développé par la société américaine Scooter Software permet de comparer des éléments : des répertoires entiers, des dossiers, mais aussi des fichiers, selon plusieurs méthodes informatiques, qui vont de la comparaison des métadonnées des items (la taille, l'horodatage, le nom et la casse des noms de fichiers et de leurs attributs) à la comparaison intégrale des contenus (via des sommes de contrôle, ou par le biais d'une comparaison binaire, octet par octet – voir Figure 22 : « Paramètres de session » pour la comparaison de dossiers, *Beyond Compare*). Ces différents niveaux de comparaison dépendent toujours des noms de dossiers ou de fichiers : « Par défaut, une session de dossiers aligne les fichiers et les sous-dossiers par nom, à gauche et à droite » écrit d'ailleurs *Beyond Compare* dans l'aide contextuelle du logiciel (Scooter Software 2022). Cela signifie donc que les paramètres indiqués (comparaison des métadonnées, ou comparaison binaire des contenus par exemple) ne s'appliquent que sur les éléments qui portent le même nom – on verra dans la partie consacrée aux études de cas qu'il est possible, et parfois utile, d'« aligner » deux éléments cote à cote afin de forcer leur comparaison.

Si la comparaison détaillée du contenu intégral de deux fichiers individuels (fichiers texte, image, ou tableur par exemple, que ce soit pixel par pixel, ou ligne à ligne) est relativement peu pertinente dans le cadre d'une évaluation macroscopique d'un fonds d'archives contenant plusieurs dizaines de milliers de documents numériques (nous verrons dans les études de cas que cette fonctionnalité peut cependant être utile sur un échantillon de fichiers bien particuliers, voir « 4.5.2 Comparer les métadonnées des fichiers »), la comparaison de deux répertoires ou de deux dossiers et de leurs contenus (par métadonnées, sommes de contrôle ou comparaison binaire) peut s'avérer efficace dans le cadre d'un traitement général des redondances (des éléments identiques ou fortement semblables). Grâce à *Beyond Compare* et aux logiciels de cette même « famille » (dont fait aussi partie *WinMerge*), il est possible d'analyser facilement la composition de répertoires (contenant des sous-dossiers) ou de

dossiers (contenant des fichiers) pour faire apparaître non seulement les éléments identiques, mais également les éléments différents (si deux fichiers qui possèdent le même intitulé ont été modifiés d'un répertoire à l'autre) et les documents « orphelins » (ceux qui ne se trouvent donc que dans un seul des répertoires et dossiers comparés).

Figure 22 : « Paramètres de session » pour la comparaison de dossiers, *Beyond Compare*



Comme nous l'expliquions précédemment, la seule particularité de ce type de logiciels (qui tient donc à leur *fonctionnement* même), c'est que l'utilisateur doit déterminer manuellement quels sont les répertoires ou les dossiers qu'il souhaite comparer. Il est donc nécessaire d'avoir une connaissance préalable de l'arborescence et du plan de classification initial du producteur des documents, pour savoir quels répertoires doivent être analysés en priorité (en d'autres termes, on ne peut pas « jeter » toutes les données dans le logiciel, à l'inverse de *AllDup* ou de *TreeSize* qui parcourent l'ensemble des fichiers avant d'afficher les résultats correspondant au(x) filtre(s) de recherche).

La principale différence qui distingue *Beyond Compare* de *WinMerge* tient à leur interface graphique et à l'affichage des comparaisons. *WinMerge* dispose tous les éléments dans un seul panneau, sous la forme d'une liste tabulaire<sup>10</sup>, avec les colonnes « Nom du fichier », « Dossier », « Résultat de la comparaison » puis les métadonnées pour chaque élément de « droite », de « gauche » et du « milieu ». L'affichage peut respecter l'arborescence (avec des dossiers et des sous-dossiers, mais toujours sous la forme d'une liste) ou présenter tous les éléments à la suite, à plat ; *WinMerge* utilise enfin toute une gamme d'icônes pour représenter visuellement les différences entre deux ou trois éléments.

<sup>10</sup> La rubrique « aide » de *WinMerge* parle effectivement de « tabular list » dans sa présentation de la comparaison de dossiers : « If you selected two folders in the Open dialog, the Folder Compare window is opened. The Folder Compare window is a tabular list of items found in the compared folders. Each row displays information about a found file, with the file name in the left column and additional information in the other columns » (*WinMerge* [sans date]).

Figure 23 : Affichage des résultats de la comparaison entre deux répertoires (« ROUSSEAU\_2011 » et « ROUSSEAU\_2012 »), Fonds Reusser, WinMerge

[illegible]

*Beyond Compare* présente quant à lui les résultats sous une forme toute différente : la fenêtre du logiciel contient effectivement deux panneaux (et non un seul comme *WinMerge*) dans lesquels sont disposés chacun des répertoires ou des dossiers que l'on est en train de confronter. Le logiciel propose également une vision structurée (respectant l'arborescence initiale, avec les différents niveaux de profondeur) ou une version sous forme de liste (le logiciel utilise d'ailleurs les formules « Ignorer la structure de dossiers » et « Aplatir » pour cet affichage particulier), avec toujours l'existence des deux panneaux, un par répertoire / dossier. Si *Beyond Compare* ne propose « que » la comparaison de deux répertoires et non de trois comme le fait *WinMerge*, l'affichage proposé le rend plus facile d'utilisation, plus convivial et surtout plus intuitif, puisqu'il respecte la structure interne à chaque répertoire – l'utilisateur retrouve sous ses yeux l'arborescence avec laquelle il a l'habitude d'interagir dans un gestionnaire de fichiers traditionnel. En outre, *Beyond Compare* indique les résultats de la comparaison non pas sous la forme d'icônes associées à chaque élément, mais de couleur, ce qui allège considérablement la vue et permet de détecter plus aisément les redondances et les dissemblances. Enfin, *Beyond Compare* offre la possibilité d'afficher de manière sélective les éléments identiques, les éléments qui ne se trouvent que dans le répertoire de gauche / de droite (nommés « orphelins ») ainsi que les éléments les plus récents selon les répertoires. Pour toutes ces raisons, *Beyond Compare* nous semble plus indiqué pour l'évaluation des supports de données que *WinMerge* et nous le recommandons donc en priorité à la Cinémathèque.



Figure 24 : Légendes des couleurs pour la comparaison de dossiers et affichages sélectifs des résultats, *Beyond Compare*

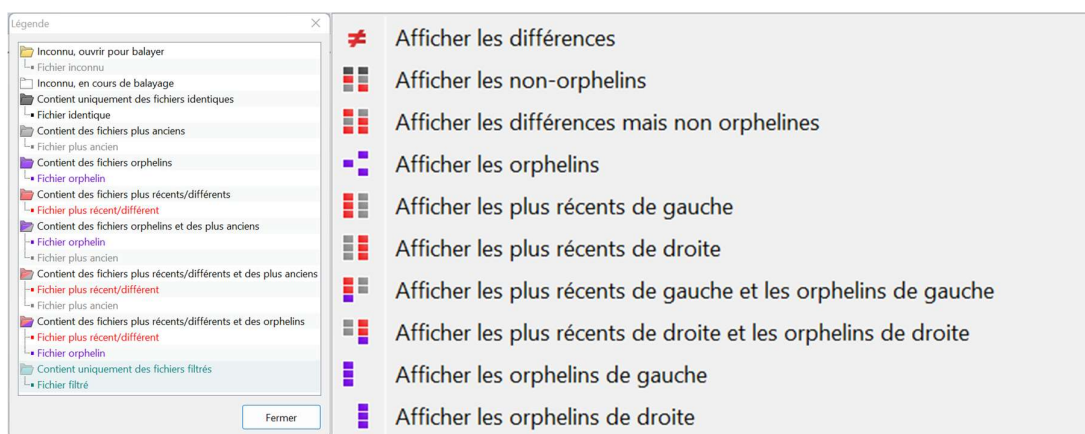
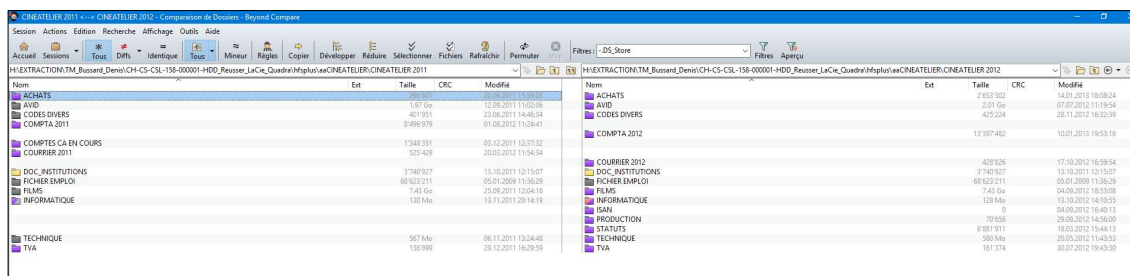


Figure 25 : Affichage des résultats de la comparaison entre deux répertoires, *Beyond Compare*, Fonds Reusser



### 3.3.4.2.2 TreeSize

On a déjà eu l'occasion de présenter les fonctionnalités proposées par *TreeSize* dans les sections précédentes. Cet outil généraliste et multi-tâches offre encore la possibilité de comparer très facilement les métadonnées des fichiers, sur la base des propriétés suivantes : « Nom, taille et date », « Nom et taille », « Nom », « Nom sans extension », « Taille et date » et « Taille ». *WinCatalog* propose à peu de choses près les mêmes catégories, dans un menu intitulé « Doublons » avec les particularités suivantes : il ne propose pas la recherche du « Nom sans extension », mais offre en revanche la possibilité de comparer les métadonnées Exif et ID3 plus spécifiques aux fichiers graphiques et audio ; les propriétés peuvent en outre être librement combinées, contrairement à *TreeSize* qui offre un menu déroulant au lieu de cases à cocher. En outre, on retrouve les différences qui distinguaient ces deux logiciels dans la recherche des redondances strictes quant à l'affichage des résultats : *TreeSize* propose des regroupements d'éléments identiques dans un seul onglet ainsi que quelques statistiques et des filtres de sélection (semi-)automatisée, tandis que *WinCatalog* ne propose qu'une liste complète de tous les éléments, sans aucun moyen de sélectionner (semi-)automatiquement des éléments, et sans offrir la moindre statistique sur les résultats, autre que le nombre total d'éléments trouvés. Si les éléments ne sont pas (forcément) identiques en termes de contenu, *TreeSize* comme *WinCatalog* intègrent cependant ces modes de détection dans l'onglet qu'ils réservent à la « Recherche de doublons » – ce qui témoigne bien de la finalité de la recherche. *TreeSize* indique d'ailleurs dans son manuel d'utilisateur à quelles fins peuvent par exemple servir ce type de filtres : « Nom et taille » pour les fichiers déplacés (« This is helpful in case

files had been moved from one location to another, which might modify this time stamp ») ; « Nom sans extension » pour les fichiers compressés (« This can be interesting in case you are searching for duplicated backup files or e.g. row-data and compact image or video files ("MyPhoto.bmp" - "MyPhoto.png ») ou encore « Taille et date » pour les copies réalisées par erreur (« Accidental copies with names such as "Copy of ..." can be identified quickly, using this method. ») (Jam Software (Joachim Marder) 2022, p. 113-114). Nous étudierons quelques cas particuliers dans le cadre de l'étude de cas (voir chapitre « 4.5.2 Comparer les métadonnées des fichiers ») en utilisant en particulier la recherche de fichiers possédant le même intitulé mais des extensions différentes, grâce aux résultats de recherche fournis par *TreeSize* dans le Fonds Reusser.

### 3.3.4.2.3 *AntiDupl*

Le logiciel *AntiDupl*, créé par deux développeurs à partir de 2003, traite uniquement les fichiers graphiques. S'il permet d'identifier les images strictement identiques, sa spécificité, et la raison pour laquelle nous avons décidé de le tester, c'est la reconnaissance d'images qui possèdent un certain degré de similarité. Pour cela, *AntiDupl* utilise deux algorithmes : « Mean square difference » (MSD, soit la différence quadratique moyenne, un algorithme qui « calculate a mean square deviation of brightness for each couple of images » (AntiDupl 2020a)) et « Index of structural similarity » (SSIM, soit l'index de similarité structurelle entre deux images, un algorithme développé à partir de 2004 qui étudie trois paramètres – la luminosité, le contraste et la structure d'une image – pour obtenir un score qui indique le degré de similarité entre deux fichiers graphiques<sup>11</sup>).

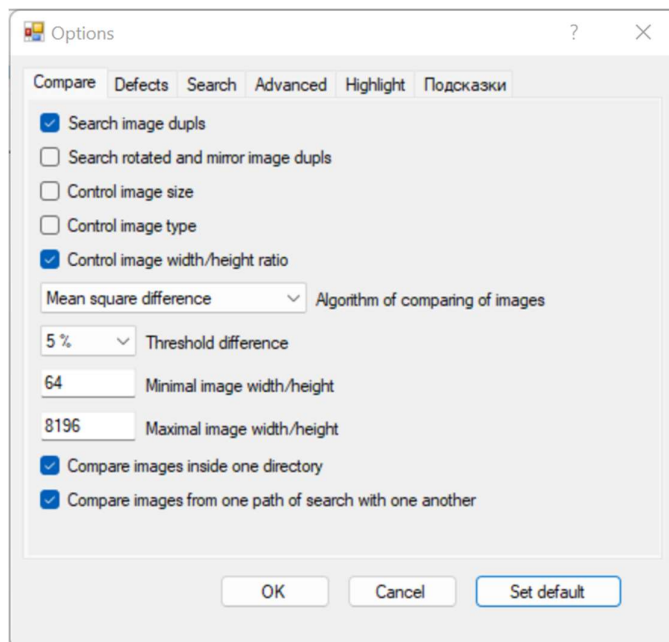
À l'instar de ce que propose *AllDup* dont il a été question plus haut, le logiciel de comparaison d'images *AntiDupl* propose également différents modes de recherche, permettant d'analyser un seul ou plusieurs répertoires de manière croisée (*inter*-répertoires), plusieurs sous-dossiers à l'intérieur du ou des répertoire(s) en question (*inter*-dossiers), ainsi que d'étudier les images qui se trouvent au sein d'un seul sous-dossier (*intra*-dossier). Le paramétrage du logiciel est cependant moins aisé et explicite que celui d'*AllDup*, et il est nécessaire de combiner le nombre de répertoires généraux (soit les répertoires sources, dans la fenêtre « path ») et les deux options proposées dans la fenêtre « Options » du logiciel : « Compare images inside one directory » et « Compare images from one path of search with one another »<sup>12</sup>.

---

<sup>11</sup> Nous n'entrerons guère dans le détail technique de l'application de tels algorithmes et on se reportera à la publication « séminale » au sujet de cet index de similarité structurelle pour plus de précision (Wang et al. 2004).

<sup>12</sup> Le manuel d'utilisateur du logiciel livre les explications suivantes à propos de ces deux options de recherche : « **Compare images inside one directory** - when this option is enabled, the program compares with each other the pictures located both in one directory and in the different. Otherwise, the program will compare among themselves only the images which aren't lying in one directory. By default, this option is disabled. » et « **Compare images from one path of search with one another** - when this option is enabled, the program compares one with another pictures that are located in the same search path. Otherwise, the program will only compare the images from different search paths. By default, this option is disabled » (AntiDupl 2020a).

Figure 26 : Menu « Options » (avec les valeurs par défaut), *AntiDupl*

















Dans le détail, en combinant les options, on obtient les huit modes de recherche suivants :

- Si l'on indique un seul chemin (« path »), les résultats de recherche sont les suivants :
  1. Décocher les deux : aucun résultat n'apparaît ;
  2. Cocher que « [...] inside one directory » : aucun résultat ne s'affiche ;
  3. Cocher que « [...] from one path of search » : comparaison des images identiques ou similaires se trouvant dans le même répertoire (chemin principal), mais dans des dossiers distincts (*inter*-dossiers)
  4. Cocher les deux : comparaison des images identiques ou similaires se trouvant dans le même répertoire (chemin principal), dans des dossiers distincts et au sein d'un même dossier (*inter* et *intra*-dossiers)
- Si l'on indique plusieurs chemins (« path »), les résultats de recherche sont les suivants :
  5. Décocher les deux : comparaison des images se trouvant uniquement dans l'un et dans l'autre (ou les autres) répertoire(s) principaux indiqués (recherche croisée, *inter*-répertoires) ;
  6. Cocher que « [...] inside one directory » : [résultats identiques aux précédents]
  7. Cocher que « [...] from one path of search » : comparaison des images identiques ou similaires se trouvant dans plusieurs répertoires (chemins indiqués) et dans des dossiers distincts du même répertoire (*inter*-répertoires et *inter*-dossiers).
  8. Cocher les deux : comparaison des images identiques ou similaires se trouvant dans plusieurs répertoires (chemins indiqués), dans des dossiers distincts du même répertoire, et au sein d'un même dossier (*inter*-répertoires, *inter* et *intra*-dossiers).

Le logiciel fonctionne extraordinairement rapidement et les résultats s'affichent alors sous la forme de deux panneaux : l'un, à gauche, avec une prévisualisation des deux images comparées (on reconnaît, sous les images, les métadonnées qui leur sont associées, avec le surlignage des différences en rouge) ; et un panneau contenant les informations détaillées relatives aux fichiers (avec des colonnes à propos du nom des fichiers, de leur chemin contenant, de leurs dimensions, de leur extension et de leur taille). Figurent en outre des informations propres à la comparaison : la *Difference* exprimée sous la forme d'un pourcentage (plus ce dernier est faible, plus les images sont similaires – voire identiques lorsque le valeur est de zéro –, et plus le chiffre est élevé, plus les images sont dissemblables) ; la *Transformation* qui indique l'opération à réaliser (rotation de l'image de 90, 180 ou 270 degrés, ou renversement horizontal / vertical si les images sont « en miroir ») pour que les deux images comparées coïncident ; enfin la colonne *Hint (Recommandation)* qui indique quelle image doit être supprimée en priorité selon le logiciel<sup>13</sup>. L'exemple ci-dessous, tiré du Fonds Reusser, représente les images similaires contenues dans le groupe numéro 432 (la paire surlignée en blanc dans le groupe est celle qui apparaît dans le panneau de visualisation) provenant d'un dossier qui contient des clichés du tournage de *Ma Nouvelle Héloïse* (nous avons alors paramétré le logiciel pour qu'il utilise l'algorithme *Mean Square Difference* avec une valeur de tolérance de 5% et le logiciel a livré 9'556 paires d'images similaires).

Figure 27 : Résultats de recherche d'images similaires, *AntiDupl*, Fonds Reusser

AntiDupl - default



File Explorer icon

Folder icon

Image icon

Print icon

Zoom in icon

Zoom out icon

Reset zoom icon

Copy icon

Paste icon

Delete icon








Undo icon

























Redo icon

Help icon

Close icon

15% | Mean up



	Type	Difference	Transformation	Hint	Name	In folder	Group	Dimensions	Image type	Size	Business	Blurring
	IMG_3418.JPG	0.00			IMG_3418.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	430	3648 x 2736	JPG	5 034 KB	0.02	0.69
	IMG_3418.JPG	0.00			IMG_3418.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	430	3648 x 2736	JPG	5 034 KB	0.02	0.69
	IMG_3418.JPG	0.00			IMG_3418.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	430	3648 x 2736	JPG	5 034 KB	0.02	0.69
	IMG_3427.JPG	0.00			IMG_3427.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	431	3648 x 2736	JPG	4 103 KB	0.29	0.72
	IMG_3427.JPG	0.00			IMG_3427.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	431	3648 x 2736	JPG	4 103 KB	0.29	0.72
	IMG_3427.JPG	0.00			IMG_3427.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	431	3648 x 2736	JPG	4 103 KB	0.29	0.72
	IMG_3427.JPG	0.00			IMG_3427.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	431	3648 x 2736	JPG	4 103 KB	0.29	0.72
	IMG_3427.JPG	0.00			IMG_3427.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	431	3648 x 2736	JPG	4 103 KB	0.29	0.72
	IMG_3427.JPG	0.00			IMG_3427.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	431	3648 x 2736	JPG	4 103 KB	0.29	0.72
	DSC_3440.JPG	2.70			DSC_3440.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3439.JPG	2.70			DSC_3439.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3441.JPG	2.70			DSC_3441.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3439.JPG	4.75			DSC_3439.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3438.JPG	4.75			DSC_3438.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3440.JPG	4.75			DSC_3440.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3441.JPG	4.94			DSC_3441.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3438.JPG	4.94			DSC_3438.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3439.JPG	4.94			DSC_3439.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3439.JPG	4.94			DSC_3439.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3440.JPG	4.94			DSC_3440.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3441.JPG	4.94			DSC_3441.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3438.JPG	4.94			DSC_3438.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3439.JPG	4.94			DSC_3439.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3440.JPG	4.94			DSC_3440.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3441.JPG	4.94			DSC_3441.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3438.JPG	4.94			DSC_3438.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3439.JPG	4.94			DSC_3439.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3439.JPG	4.94			DSC_3439.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3440.JPG	4.94			DSC_3440.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3441.JPG	4.94			DSC_3441.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3438.JPG	4.94			DSC_3438.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3439.JPG	4.94			DSC_3439.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3440.JPG	4.94			DSC_3440.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3441.JPG	4.94			DSC_3441.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3438.JPG	4.94			DSC_3438.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3439.JPG	4.94			DSC_3439.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3439.JPG	4.94			DSC_3439.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3440.JPG	4.94			DSC_3440.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3441.JPG	4.94			DSC_3441.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3438.JPG	4.94			DSC_3438.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3439.JPG	4.94			DSC_3439.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3440.JPG	4.94			DSC_3440.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3441.JPG	4.94			DSC_3441.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3438.JPG	4.94			DSC_3438.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3439.JPG	4.94			DSC_3439.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3439.JPG	4.94			DSC_3439.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3440.JPG	4.94			DSC_3440.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3441.JPG	4.94			DSC_3441.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3438.JPG	4.94			DSC_3438.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3439.JPG	4.94			DSC_3439.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3440.JPG	4.94			DSC_3440.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3441.JPG	4.94			DSC_3441.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3438.JPG	4.94			DSC_3438.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3439.JPG	4.94			DSC_3439.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3439.JPG	4.94			DSC_3439.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3440.JPG	4.94			DSC_3440.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3441.JPG	4.94			DSC_3441.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3438.JPG	4.94			DSC_3438.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3439.JPG	4.94			DSC_3439.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3440.JPG	4.94			DSC_3440.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3441.JPG	4.94			DSC_3441.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3438.JPG	4.94			DSC_3438.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3439.JPG	4.94			DSC_3439.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3439.JPG	4.94			DSC_3439.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3440.JPG	4.94			DSC_3440.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3441.JPG	4.94			DSC_3441.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3438.JPG	4.94			DSC_3438.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3439.JPG	4.94			DSC_3439.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3440.JPG	4.94			DSC_3440.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3441.JPG	4.94			DSC_3441.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3438.JPG	4.94			DSC_3438.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34
	DSC_3439.JPG	4.94			DSC_3439.JPG	H:\EXTRACTION\TM_Bussard_Denis\CH-CS-CSL-158-00001-HDD_Reusser_LaCie	432	3872 x 2592	JPG	8 919 KB	21.37	1.34

774 KB

3872 x 2592

22.26

1.75

JPG

EXP

H:\EXTRACTION\TM\_Bussard\_Denis\CH-

846 KB

3872 x 2592

20.23

1.88

JPG

EXP

H:\EXTRACTION\TM\_Bussard\_Denis\CH-

Total: 9556

Current: 1778

Selected: 10

En raison du fonctionnement interne du logiciel (algorithme de comparaison entre deux fichiers graphiques), les résultats de recherche peuvent être très élevés. En effet, le logiciel comparant les fichiers un à un, image par image, les résultats affichés concernent toujours deux éléments seulement et peuvent ainsi être très nombreux : pour trois images qui se ressemblent, on peut donc avoir, au maximum, un résultat de comparaison pour les images 1 et 2, 1 et 3 et 2 et 3 (et donc 3 résultats pour 3 images, 6 résultats pour 4 images, 10 résultats pour 5 images, et ainsi de suite, avec pour  $n$  images, un nombre de résultats qui équivaut à la somme des entiers de  $n-1$ ). Dans l'exemple ci-dessus, si on regarde attentivement le nom des photographies, on constate qu'il y a « seulement » six clichés différents, tous stockés dans le même répertoire et que le logiciel livre onze paires de comparaison (le nombre maximal aurait donc été de 15 résultats, mais il faut croire que la comparaison de certaines images dépassait le seuil de tolérance choisi, de 5 %, pour être reconnues comme similaires). Si le nombre de résultats peut paraître ainsi disproportionné et difficilement abordable, le logiciel propose des regroupements de paires qui ont obtenu des résultats de comparaison proches, ce qui peut s'avérer extrêmement utile pour traiter ensemble des paires d'images représentant quasiment le même « objet » (cette information figure dans la colonne *Group*<sup>14</sup>, que l'on peut trier pour faire apparaître à la suite les paires d'images similaires). Une précision s'impose ici : il est parfois nécessaire d'augmenter le seuil de tolérance des différences (la valeur, en pourcentage, à partir de laquelle le logiciel affiche le résultat de la comparaison entre deux images) pour faire apparaître dans un seul groupe tous les membres d'un même « lot » de photographies qui représentent le même sujet.

Le fonctionnement des algorithmes utilisés par *AntiDupl* rend donc quelque peu difficile l'établissement de statistiques à propos des images similaires au sein du répertoire étudié (parce que le nombre de résultats ne correspond pas au nombre de fichiers et parce que les résultats varient énormément en fonction du seuil de tolérance des différences que l'on choisit). Le logiciel présente également les quelques inconvénients suivants : l'annulation ne prend en compte qu'un nombre limité d'actions précédemment réalisées (par défaut, il est possible d'annuler les dix dernières opérations réalisées au sein du programme et ce nombre peut monter jusqu'à seize seulement) ; *AntiDupl* ne prend en charge qu'un nombre limité de formats d'images (BMP, GIF, JPEG, PNG, TIFF, EMF, WMF, EXIF, ICON, JP2, PSD, DDS, TGA – il ne prend donc pas en charge le format .NEF par exemple, alors que le Fonds Reusser contient plus de 4'000 fichiers de ce format *Nikon*) ; enfin, l'export des résultats n'est pas de bonne qualité (le logiciel propose seulement de faire un copier-coller manuel des résultats de recherche).

En dépit de ces inconvénients, *AntiDupl* possède de nombreux arguments en faveur de son utilisation par la Cinémathèque (outre le fait qu'il soit facile d'utilisation et qu'il ne nécessite pas de compétences informatiques avancées) :

- Le regroupement des paires d'images similaires (via la colonne *Group*) pour travailler par lot de photographies ;
- Le paramétrage manuel du seuil de tolérance des différences, permettant de traiter les images identiques dans un premier temps, puis les images qui possèdent peu de différences, et ainsi de suite ;

---

<sup>14</sup> *AntiDupl* définit ainsi les informations regroupées dans la colonne intitulée « Group » : « searched pairs of duplicate images can be combined into groups with similar results. This column displays the number of such groups » (*AntiDupl* 2020c)

- L'affichage des photographies directement voisines à celles que nous sommes en train de comparer, permettant de prendre connaissance du contexte (très direct ici) dans lequel les images sont enregistrées.

#### 3.3.4.3 Comparaison de données : conclusion

À ce stade de la recherche, il est difficile de tirer des conclusions générales à propos de la comparaison de données dans le domaine archivistique, tant le domaine est complexe et en pleine expansion (comme en témoignent par exemple les travaux menés par le laboratoire « Humanités digitales » de l'EPFL autour du projet intitulé *Replica* qui visait à permettre la recherche d'attributs visuels – des formes ou des motifs similaires – au sein des collections numérisées d'œuvres d'art des grands musées mondiaux (di Lenardo, Seguin, Kaplan 2016; Seguin et al. 2016)). Notons seulement que le projet de l'EPFL utilise des algorithmes d'intelligence artificielle (des réseaux neuronaux plus précisément) qui dépassent largement le cadre de cette étude. Très prometteurs, ces algorithmes ne sont pas encore à la portée de tous les archivistes ou spécialistes du patrimoine culturel, raison pour laquelle nous avons présenté ci-dessus seulement quelques logiciels, dont *AllDup* et *AntiDupl*, qui utilisent des algorithmes de comparaison d'images qui peuvent être utilisés dès aujourd'hui par un public non averti. Les développements informatiques futurs dans ce domaine doivent donc être suivis de près, pour guetter l'apparition d'un outil grand public, possédant une interface graphique facilement utilisable. D'ici-là, des outils comme ceux présentés ci-dessus (*Beyond Compare*, *TreeSize* et la paire *AllDup/AntiDupl*) permettent déjà de traiter les dossiers et les fichiers similaires à plusieurs niveaux, comme nous le verrons dans les études de cas ci-dessous (voir « 4.5 Comparaison de données »)

## 3.4 Discussion méthodologique

### 3.4.1 Formulation et abstraction

La principale difficulté méthodologique réside dans la bonne formulation des micro-fonctionnalités et informations proposées et dans le degré d'abstraction, ou de granularité, que l'on adopte. Un niveau trop élevé aurait pour conséquence de réduire à quelques intitulés vagues l'ensemble des micro-fonctionnalités offertes et de niveler ainsi ce qui peut faire la force d'un logiciel. À l'inverse, une granularité trop fine (en restant par exemple fidèle aux formulations proposées par les développeurs dans les manuels d'utilisation des logiciels) gonflerait excessivement la liste des informations et rendrait impossible les comparaisons inter-outils. Nous avons alors essayé d'adopter un point de vue intermédiaire : une conceptualisation suffisamment élevée pour permettre les comparaisons, mais assez précise pour rendre compte des spécificités de chaque outil – tout en étant attentif au fait que la formulation choisie est souvent dépendante de la dénomination proposée par le premier logiciel testé à proposer cette micro-fonctionnalité.

### 3.4.2 Autonomie et dépendance des micro-fonctionnalités et informations

Parmi les micro-fonctionnalités et informations retenues, toutes n'ont pas le même degré d'autonomie et il arrive fréquemment que plusieurs éléments soient (inter-)dépendants. Cette situation se présente notamment lorsqu'une micro-fonctionnalité d'un niveau élevé (primaire) conditionne l'existence et la bonne exécution de micro-fonctionnalités secondaires. C'est par exemple le cas pour la recherche de redondances. Si l'outil travaille uniquement via l'extraction et la comparaison des métadonnées dans un tableur, les micro-fonctionnalités suivantes, liées à l'interface graphique du logiciel, ne sont pas remplies ; il en va de même pour l'affichage des résultats par liste (« à plat ») ou par groupe (par « onglet ») puisque de nombreuses informations quantitatives relatives aux groupes de redondances ne sont pas délivrées par les logiciels affichant les résultats uniquement sous forme de liste.

Pour ce qui concerne les informations, on rencontre surtout cette difficulté avec les « champs calculés », c'est-à-dire lorsqu'une information est dérivée d'autres valeurs quantitatives proposées par l'outil. La totalité des proportions (exprimées en pourcentage) du nombre et du poids des dossiers et fichiers vis-à-vis des répertoires parents dépendent ainsi d'autres valeurs quantitatives ; de même pour les sommes totales d'éléments (dossiers et fichiers par exemple) qui sont calculées grâce à des informations primaires délivrées par le logiciel. À moins que ce « champ calculé » soit d'une grande utilité pour l'évaluation, il a été décidé, afin d'atténuer ce biais, d'attribuer un coefficient peu élevé (1 ou 2) à ces informations « dérivées ».

### 3.4.3 Accomplissement intégral ou partiel d'une fonctionnalité

L'attribution d'une valeur positive (1) ou nulle (0) à un logiciel pour une catégorie pose également des problèmes de granularité : comment distinguer avec un codage binaire l'accomplissement total ou partiel d'une micro-fonctionnalité ? Il arrive en effet fréquemment que deux outils proposent des fonctionnalités similaires en termes conceptuels, mais que leur réalisation pratique diffère. Prenons par exemple le cas du « Développement par niveau de l'arborescence ». Si *TreeSize* offre la possibilité de développer *toute* l'arborescence au niveau 1, 2, 3, 4 ou 5, *Droid* propose quant à lui de développer les trois seuls niveaux qui sont directement subordonnés à un dossier particulier (on peut donc développer les trois premiers niveaux de toute l'arborescence, ou les trois niveaux qui suivent directement une branche du répertoire). Le même problème se pose avec ce que nous avons appelé les « Statistiques

visuelles des formats – surlignage dans l'arborescence » : *TreeSize* indique dans le gestionnaire au complet le nombre de fichiers d'une extension donnée contenus dans chaque dossier (quitte à indiquer 0 si le dossier n'en comporte pas), tandis que *Droid* masque, dans le gestionnaire de fichiers, les dossiers ne comportant aucun élément d'une extension précise et n'indique pas combien de fichiers du format sélectionné sont contenus dans les dossiers visibles. Si l'affichage choisi par *TreeSize* est plus riche (quantitativement et « contextuellement »), il génère beaucoup de « bruit », à l'inverse de *Droid*, qui perd en contexte ce qu'il gagne en clarté.

Les arbitrages sont donc parfois difficiles et le codage binaire perd en subtilité ce qu'il gagne en efficacité... À ce stade de la recherche, nous avons décidé d'attribuer une valeur positive (1) à un logiciel s'il proposait la micro-fonctionnalité, quelle que soit la manière dont il l'accomplissait – les subtilités d'application sont détaillées dans les résultats et conclusions. Un codage plus fin serait néanmoins envisageable dans un second temps pour atténuer ce biais méthodologique.



## 4. Méthodologie de tri archivistique

Au terme de cette analyse des fonctionnalités informatiques, nous aimerions proposer une esquisse de méthodologie en se basant sur les études de cas réalisées grâce aux Fonds Reusser et Simon par le biais, entre autres, des outils sélectionnés pour leur pertinence et leur utilité lors de l'évaluation d'un fonds d'archives numériques privées. Si les tâches archivistiques proposées par les *workflows* déjà publiés (voir à ce sujet la revue de littérature au chapitre 1.3.1) nous avaient permis dans un premier temps de sélectionner les fonctionnalités logicielles à étudier, à l'inverse, et dans un mouvement constant d'aller-retour et d'itérations, le test des fonctionnalités logicielles et des outils nous permet de préciser quelque peu les tâches à accomplir. Car si les termes sont parfois identiques entre une fonctionnalité et une tâche archivistique, et que les deux entretiennent un lien très étroit, elles ne se recoupent pas entièrement. À titre d'exemple, le dédoublonnage est une fonctionnalité logicielle (elle est présentée comme telle dans les listes de *features* qui figurent sur les sites internet des développeurs) et une tâche archivistique : elles se chevauchent donc mais elles ne se confondent pas. Une fonctionnalité n'est en effet rien de plus qu'un « outil intégré à un logiciel, à un site Web, à un périphérique ou à un appareil, qui permet à l'utilisateur d'effectuer une tâche ou une action » (Grand dictionnaire terminologique (GDT) 2012). Elle ne dit donc ni le comment, ni le pourquoi (et encore moins sur quels éléments les traitements proposés doivent porter). La tâche archivistique est autrement plus large et complexe, nécessitant des analyses, des décisions et des traitements qui utilisent les fonctionnalités informatiques mais qui les dépassent largement.

En termes méthodologiques, nous proposons donc de reprendre les quatre intitulés des fonctionnalités étudiées jusque-là et de décomposer ces grandes catégories en une suite de tâches et de sous-tâches, mais également de proposer des points d'analyse ou des points d'attention, qui pourront faciliter la prise de décision quant au sort final des documents. Les exemples porteront alors principalement sur les archives de Francis Reusser, autrement plus nombreuses que celles d'Ana Simon. En outre, sachant que le logiciel *Ingest Manager* de la Cinémathèque n'est pas encore opérationnel et que nous n'avons pas la possibilité de réaliser des éliminations de fichiers système ou liés aux applications, les analyses quantitatives ci-dessous portent sur le fonds entier, non expurgé (les étapes de curation du *Pré-Ingest* ne sont donc pas réalisées) – il s'agit donc bien d'exemples de types possibles d'analyse.

Enfin, la méthodologie suit une chronologie précise, et les étapes sont ordonnées ici dans l'ordre dans lequel il convient de les réaliser : du récolement à la comparaison de données, en passant par l'analyse de l'arborescence, le traitement des dossiers vides et le traitement des redondances strictes. Les tâches, sous-tâches, méthodes et instruments, points d'analyse et logiciels utilisés sont résumés dans des tableaux synthétiques par étape :

1. Tableau 12 : Extraction de métadonnées / Récolement, tâches
2. Tableau 19 : Arborescence et volumétrie, tâches
3. Tableau 20 : Traitement des dossiers vides, tâches
4. Tableau 27 : Traitement des redondances strictes, tâches
5. Tableau 29 : Comparaison de données, tâches

## 4.1 Extraction de métadonnées / Récolement

La première étape que nous recommandons pour procéder à l'évaluation et au tri archivistiques des supports de données consiste en une extraction des métadonnées nécessaires, c'est-à-dire dans l'établissement d'un récolement de qualité. Il n'est pas question de procéder à des études de cas détaillées à propos des Fonds Reusser et Simon (une comparaison de la qualité des métadonnées proposées par chaque logiciel dépasse effectivement le cadre de cette étude et pourrait faire l'objet d'une recherche à part) mais seulement de résumer sous la forme d'une méthodologie applicable les principales tâches que cela implique.

Comme on peut le voir dans le tableau ci-dessous, il s'agit d'un processus fait de quatre grandes étapes : la sélection des métadonnées ; l'extraction des métadonnées ; la compilation des métadonnées et enfin la sauvegarde du récolement ainsi généré dans le dossier d'acquisition du Fonds. L'analyse des logiciels, par des études de cas et en parcourant la documentation qu'ils proposent sur leur site internet, doit permettre de dresser la liste des métadonnées que l'on peut extraire ; un intense travail suit cette étape d'identification : il s'agit de comparer les champs et de sélectionner les informations importantes. Nous proposons de sélectionner au minimum les métadonnées auxquelles nous avons attribué un coefficient 3, listées dans le Tableau 1 : « Extraction de métadonnées / Récolement », par métadonnées / informations, résumé. Cela concerne donc les outils suivants : *DROID*, *Karen's Directory Printer* et *TreeSize Professional*. Une fois la sélection et l'extraction réalisées, il convient de lier les trois fichiers par le biais d'une clé relationnelle ou de plusieurs clés combinées et de générer un seul fichier contenant toutes les informations. Toutes les étapes à l'exception de l'extraction elle-même peuvent être réalisées à l'aide d'un tableur (les tests de logiciels ont été effectués avec *Excel* et des fichiers « .csv » ou « .tsv »). Le récolement final sera alors conservé, idéalement sous la forme d'un fichier non-propritaire (« .csv ») et d'un PDF-A, dans le dossier d'acquisition du fonds. Cet instantané fidèle des données qui ont été remises à l'institution patrimoniale servira pour les analyses suivantes (de l'arborescence, des redondances ou de la comparaison de données) ; mais il constituera également le document probant qui atteste de la bonne remise des fichiers *par* et *auprès* des donateurs (un exemplaire pourra être remis aux producteurs des documents ou à leurs ayants droit) ; une copie, sous la forme d'une version non modifiable (le PDF-A par exemple), pourra enfin être remise, moyennant l'autorisation des ayants droit, aux chercheurs et chercheuses qui en feraient la demande pour étudier le support de données tel qu'il a été remis à la Cinémathèque (s'ils estiment nécessaires de remonter à la source en étudiant un disque dur avant traitement).

Tableau 12 : Extraction de métadonnées / Récolement, tâches

Tâche	Sous-tâches	Méthodes et instruments	Points d'analyse	Traitement	Outils
Extraction des métadonnées Récolement	Sélection des métadonnées	* Mode d'emploi des logiciels * Études de cas	* Identification des métadonnées extraites * Comparaison des métadonnées par logiciel * Sélection des métadonnées nécessaires	Cartographie (mapping) des métadonnées	* <i>DROID</i> * <i>Karen's Directory Printer</i> * <i>TreeSize Professional</i> * <i>Tableur</i>
	Extraction des métadonnées	Interface des logiciels	* Identification des noms de champs * Contrôle de la qualité de l'extraction	Enregistrement des fichiers en format ouvert	* <i>DROID</i> * <i>Karen's Directory Printer</i> * <i>TreeSize Professional</i> * <i>Tableur</i>
	Compilation des métadonnées	Base de données relationnelle (avec clé relationnelle)	-	* Normalisation des intitulés de colonnes * Création d'une clé relationnelle * Création d'un récolement unique	* <i>Tableur</i>
	Sauvegarde du récolement initial	-	-	Enregistrement dans le dossier d'acquisition	* <i>Tableur</i> * <i>PDF-A</i>

## 4.2 Analyse de l'arborescence

Grâce aux logiciels *Archifiltre*, *TreeSize Professional* et *WinDirStat*, mais aussi par l'étude des métadonnées extraites précédemment, il est possible d'analyser l'arborescence initiale du support de données. Il s'agira entre autres de s'interroger sur la manière dont les documents sont classés, à quel niveau de profondeur, et quelle est la structure générale des répertoires. Cela doit permettre à l'archiviste de prendre connaissance du Fonds, comme on le ferait lors d'une analyse préliminaire de documents analogiques avec le donateur ou ses ayants droit en observant le matériel original de conservation (les boîtes, les cartons), les dossiers contenant (les classeurs, les fourres en plastique annotées, etc.) ou l'organisation physique et matérielle des archives (les pièces d'un appartement, les rayonnages d'une bibliothèque, les tiroirs d'un bureau). Si l'arborescence peut être considérée comme une représentation mentale du créateur des documents, il vaut la peine de s'y attarder avant de procéder au traitement proprement dit – d'autant plus que les traitements en question devront prendre appui sur la connaissance générale du fonds acquise lors de cette étape (on pense en particulier au traitement des redondances). Les quelques données ci-dessous donnent une première idée de la masse totale des éléments enregistrés sur les supports de données<sup>15</sup> ; reste donc à comprendre de quelle manière ils sont organisés, objet de l'étude à venir.

Tableau 13 : Arborescence et volumétrie : cas pratiques, résumé

Variables	Fonds Francis Reusser				Fonds Ana Simon			
	Archifiltre	DROID	TSP	WinDirStat	Archifiltre	DROID	TSP	WinDirStat
Nombre total d'éléments	40044	53820	42654	41678	1334	1403	1397	1397
Nombre total de dossiers	2455	4788	3948	3456	144	203	202	202
Nombre total de fichiers	37589	49032	38706	38222	1190	1200	1195	1195
Poids total du répertoire (en Go)	240.1	-	223.6	223,5	98	97,99	91.3	91.3
Niveaux de profondeur	14	x	x	x	10	x	x	x

### 4.2.1 Analyse de la profondeur de l'arborescence

Les méthodologies existantes proposent fréquemment de traiter les fonds d'archives numériques privées selon une approche « descendante », c'est-à-dire de commencer par les dossiers sommitaux, afin de prendre des décisions quant au sort final des archives au niveau des séries et des dossiers et non au niveau des fichiers individuels. Si cette approche peut s'avérer tout à fait pertinente face à la masse de données, il convient de déterminer à quelle profondeur l'analyse doit être effectuée, c'est-à-dire à quel étage de l'arborescence et à quel répertoire doit-on accorder notre attention. Pour déterminer le niveau adéquat, il est nécessaire de connaître la profondeur générale de l'arborescence, mais aussi à quel étage se trouve la majorité des dossiers et des fichiers. Nous proposons donc ci-dessous quelques approches statistiques pour comprendre quelle est la forme générale de l'arborescence et utilisons les visualisations proposées par *Archifiltre* pour repérer les niveaux de profondeur distincts des répertoires.

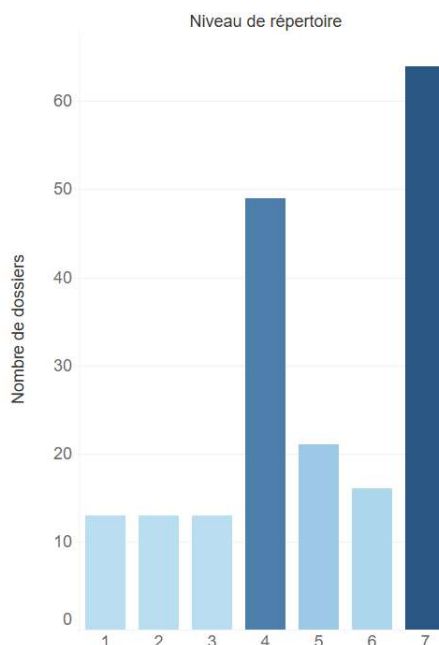
#### 4.2.1.1 Ana Simon

Les clés USB du Fonds Ana Simon contiennent jusqu'à sept niveaux de répertoire. C'est d'ailleurs à ce niveau-là que se trouve le plus grand nombre de dossiers avec 64 dossiers (presque 5 par clé) contre 49 (soit un peu moins de 4 par clé) pour le niveau 4. Si tous les dossiers de septième niveau se trouvent par définition au fond de l'arborescence, 35 dossiers

<sup>15</sup> Les valeurs en noir sont celles fournies directement par les logiciels ; en rouge, il s'agit des valeurs calculées sur la base des données récoltées.

de quatrième niveau (sur 49, soit 70 %) sont des dossiers finaux – ce qui signifie que toute l'arborescence qui se déploie sous le quatrième niveau (entre le 5<sup>e</sup> et le 7<sup>e</sup>) est comprise dans les 14 dossiers de niveau 4 restants. On ne compte en revanche que 13 dossiers pour chacun des trois premiers niveaux, soit un par clé USB. Ces informations permettent d'avoir une idée de la forme générale de l'arborescence : une première pyramide avec une structure cylindrique sur les trois premiers niveaux et un embranchement multiple au quatrième niveau ; puis une seconde pyramide, dès le quatrième niveau, qui se déploie sous une partie seulement (30 %) de la première structure pyramidale (si on regarde dans le détail, cela concerne 8 clés sur 13 : n° 3, 4, 5, 7, 8, 9, 10 et 11).

Figure 28 : Nombre de dossiers par niveau de répertoire, Fonds Simon

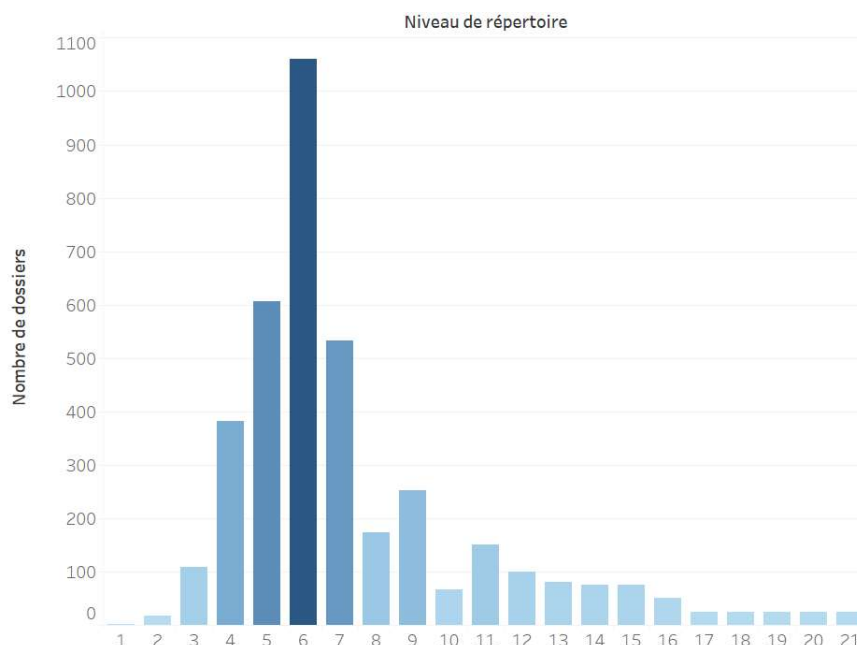


#### 4.2.1.2 Francis Reusser

Le disque dur externe de Francis Reusser contient en tout vingt-et-un niveaux de profondeur, comme on peut le voir sur la « Figure 29 : Nombre de dossiers par niveau de répertoire, Fonds Reusser ». Ces chiffres doivent toutefois être analysés en détail pour prendre leur pleine signification. En effet, alors que le premier niveau (« hsfplus ») correspond au disque dur lui-même, les dossiers se trouvant à partir du douzième niveau (entre le 12<sup>e</sup> et le 21<sup>e</sup>) ne sont guère significatifs ni intéressants pour comprendre l'arborescence générale puisqu'il s'agit d'environ 500 éléments organisés le plus souvent de manière cylindrique à propos de matériel informatique (AVID et HP Laser jet)<sup>16</sup>. L'arborescence est donc surtout concentrée entre les niveaux 2 et 11, avec un nombre de dossiers qui croît de manière constante (adoptant une forme de pyramide aux arêtes régulières) entre le premier et le sixième niveau, qui comporte le plus d'éléments (1'060, soit 27 % de tous les dossiers). Après l'expansion continue de l'arborescence jusqu'au sixième niveau, seuls quelques parties de la structure générale font donc l'objet de subdivisions plus fines.

<sup>16</sup> Classés dans les dossiers de niveau 1 à 3 suivants : « hfsplus\aaCINEATELIER\CINEATELIER » de 2009 à 2013, ils font ensuite partie des embranchements relatifs à « AVID » ainsi qu'à « INFORMATIQUE\HP\_doc Francais\Manuals\hp\_LaserJet\_1320 » et suivants.

Figure 29 : Nombre de dossiers par niveau de répertoire, Fonds Reusser



Ces conclusions intermédiaires se confirment lorsqu'on analyse l'emplacement des *dossiers finaux non vides* : l'arborescence se termine majoritairement (dans plus de 60 % des cas) aux niveaux 5 à 7, tandis qu'il n'existe aucun dossier final aux profondeurs 17 à 20 (ce qui indique bel et bien l'existence de cylindres). Il est enfin intéressant de regarder où se trouvent globalement les fichiers au sein de cette arborescence – même si la structure générale s'étudie prioritairement au niveau des dossiers, il peut être utile de confirmer les hypothèses de travail via une étude de la répartition des fichiers par niveau de profondeur (car une arborescence peut compter de nombreux dossiers comportant peu de fichiers ou, à l'inverse, un nombre restreint de répertoires contenant un grand nombre d'items).

Figure 30 : Nombre de dossiers finaux non vides par niveau, Fonds Reusser

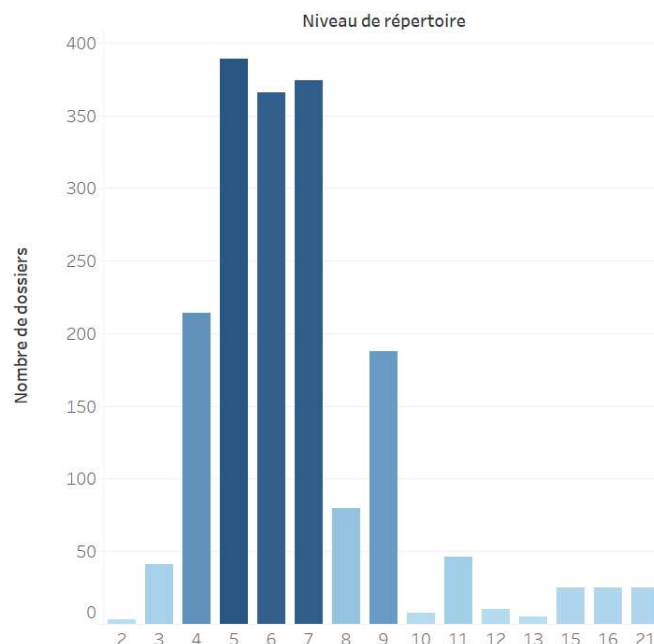
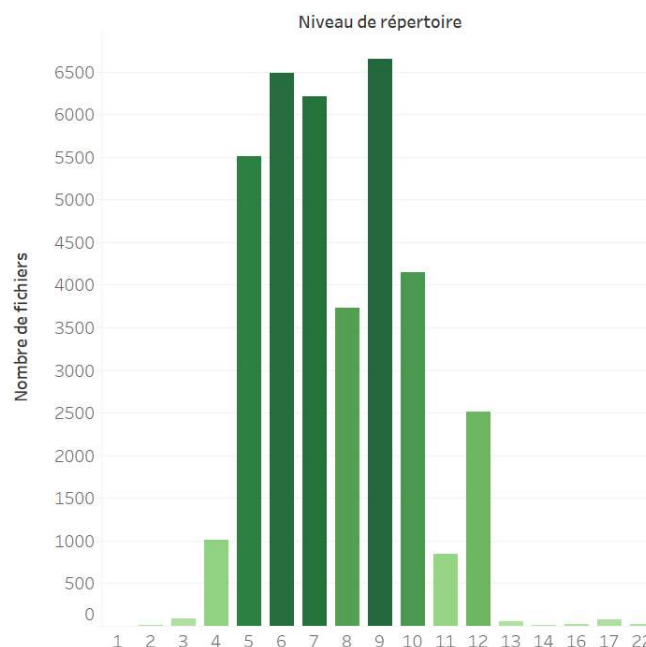


Figure 31 : Nombre total de fichiers par niveau, Fonds Reusser



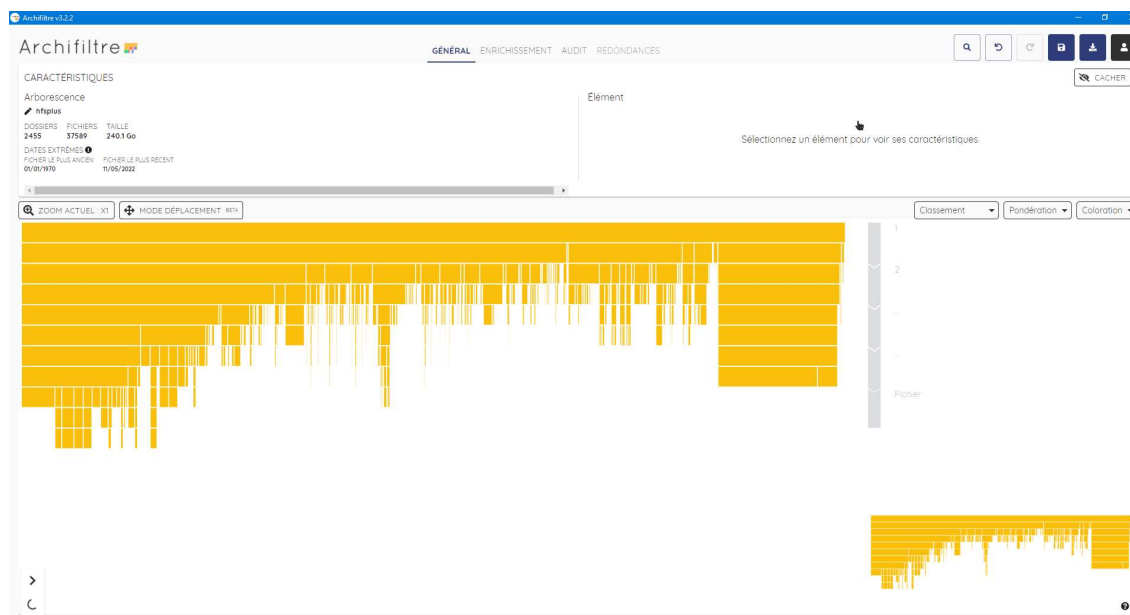
On constate alors que l'immense majorité des fichiers du Fonds Reusser se trouvent entre les niveaux 5 et 10, ce qui correspond globalement à la distribution des *dossiers finaux non vides* par profondeur étudiée ci-dessus. On gardera toutefois en tête le lien de subordination qui existe entre les dossiers et les fichiers, puisque les fichiers se trouvant par exemple au sixième niveau de l'arborescence sont contenus dans les dossiers *finaux* et *non finaux* de cinquième niveau – il s'agit dans ce dernier cas de fichiers qui se trouvent à la *racine* au sixième niveau. En d'autres termes, une partie seulement des fichiers d'un niveau ( $n$ ) sont contenus dans les dossiers finaux du niveau supérieur (niveau  $n+1$ ), les autres sont des fichiers qui se trouvent à la racine d'une arborescence qui se prolonge (voir le chapitre 4.2.2.2 pour un exemple d'analyse des fichiers à la racine dans le Fonds Reusser).

Le graphique concernant les fichiers fait cependant apparaître un cas particulier : il existe effectivement un très grand nombre d'éléments se trouvant au niveau 9 (6'645 en tout) alors que le graphique précédent (Figure 30 : Nombre de dossiers finaux non vides par niveau, Fonds Reusser) indique un nombre total peu élevé de dossiers finaux au niveau 8. Cette situation intervient dans les deux cas de figure suivants : la présence massive de fichiers à la racine (au niveau  $n$ ) ou l'existence de quelques dossiers (niveau  $n+1$ ) contenant l'immense majorité des fichiers de niveau  $n$  (c'est le cas qui se présente ici, avec 134 dossiers parents différents, pour 6'645 fichiers, qui sont d'ailleurs majoritairement des fichiers système ou en lien avec l'achat de matériel informatique par Francis Reusser).

Ces données statistiques ne sont toutefois qu'une partie de l'analyse préalable de l'arborescence en termes de profondeur. En effet, si ces données quantitatives permettent de se faire une idée générale du Fonds, elles nivellent (ou pire : elles masquent) les disparités individuelles qui existent entre les branches de l'arborescence. La visualisation proposée par *Archifiltre* permet alors de savoir si l'arborescence présente des niveaux de profondeur homogènes ou hétérogènes entre les différents répertoires qu'elle contient. À titre d'exemple, l'arborescence du Fonds Reusser se déploie sur plusieurs niveaux très hétérogènes d'un répertoire à l'autre, comme on peut le voir sur la figure ci-dessous (on le voit également sur la

Figure 13 : Onglet « Général » d'Archifiltre, classement et pondération par volume). Le répertoire de niveau 2 intitulé « aaCINEATELIER » (le premier du deuxième niveau depuis la gauche) contient non seulement le plus grand nombre d'éléments (la largeur du rectangle en témoigne, avec 24'887 fichiers), mais aussi la structure la plus profonde (10 niveaux sont visibles ici), tout l'inverse par exemple du répertoire de même niveau 2 (« UTILE ») qui contient peu d'éléments (210 fichiers) et peu de niveaux de profondeur (avec le zoom initial, seul le répertoire de niveau 2 « UTILE » apparaît – il s'agit du troisième rectangle de deuxième niveau depuis la droite, à côté de l'empilement cylindrique que forment les dossiers système).

Figure 32 : Onglet « Général », *Archifiltre*, classement par volume et pondération par nombre, Fonds Reusser



## 4.2.2 Analyse de la structure de l'arborescence

Il s'agit ensuite d'analyser la manière dont l'arborescence est globalement structurée, c'est-à-dire de comprendre comment elle est construite, quelle est son architecture générale. Pour réaliser cette tâche, c'est aux dossiers que l'on portera notre attention, via des outils de visualisation de l'arborescence (*Archifiltre*) et par le biais d'une analyse statistique. Cette dernière pourra être menée par deux moyens distincts : via la recherche avancée proposée par *TreeSize*, qui permet de paramétrer des filtres de recherche et d'exporter les résultats, ou par une extraction de métadonnées fournissant le nombre de sous-dossiers et de fichiers *par dossier individuel* et par leur traitement dans un tableur de type *Excel* – les données ci-dessous proviennent du récolement proposé par *Karen's Directory Printer*. Neuf paramètres peuvent faire l'objet d'une étude (voir la colonne « Points d'analyse » du « Tableau 14 : Analyse de la structure de l'arborescence, Fonds Simon »).

### 4.2.2.1 Ana Simon

La visualisation de l'arborescence complète proposée par *Archifiltre* permet d'identifier très rapidement la structure générale du Fonds : sur les figures ci-dessous, qui comprennent tous les supports ainsi que les images disque produites par la Cinémathèque suisse (soit les fichiers générés par *Aaru*, qui apparaissent en gris) sous une cote unique (CH-CS-CSL-136), on reconnaît sans peine les 13 clés USB sous la forme de « silos » verticaux. Ces « silos », un par clé USB, sont composés majoritairement de dossiers cylindriques qui comprennent un seul

sous-dossier. Cette organisation d'apparence très verticale laisse en réalité transparaître une organisation peu hiérarchisée, avec un grand nombre d'éléments subordonnés en bout de chaîne, disposés côte à côte. Hormis le silo constitué par les premiers niveaux de l'arborescence, on est donc plutôt face à une organisation de type horizontal. Si *Archifiltre* permet de prendre connaissance de la structure générale d'un seul coup d'œil, le logiciel se révèle en revanche assez peu utile pour la suite de l'analyse d'une telle structure (en agrandissant les éléments, on voit seulement, en bout de chaîne, un alignement de rectangles colorés représentant les différents fichiers). On constate encore sur la Figure 33 (avec une pondération par volume) l'existence de quelques fichiers volumineux de type vidéo (« .mov » ou « .mp4 », en violet) et la prédominance de sept clés sur treize ; et l'on remarque sur la Figure 34 (avec une pondération par nombre) trois groupes de clés USB distincts : quatre clés contenant peu de fichiers (n° 2, 6, 12 et 13), la clé n° 10 comportant un très grand nombre de fichiers, et un groupe intermédiaire composé de huit clés avec un nombre moyen de fichiers.

Figure 33 : Onglet « Général », *Archifiltre*, classement et pondération par volume, Fonds Simon

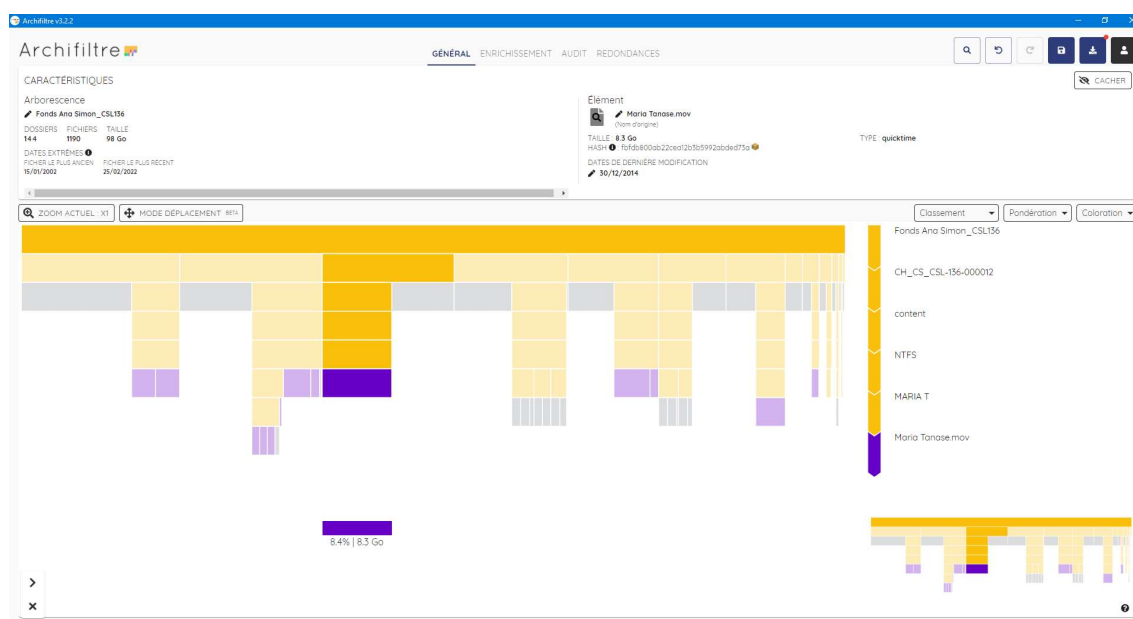
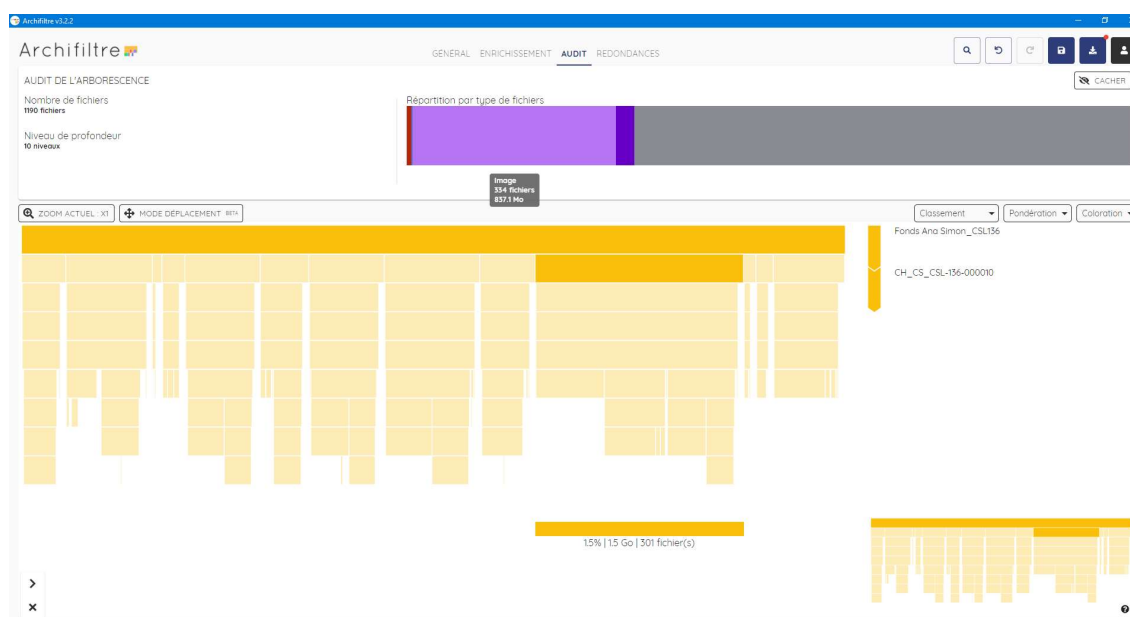




Figure 34 : Onglet « Audit », *Archifiltre*, classement par volume et pondération par nombre, Fonds Simon



Les analyses statistiques (voir Tableau 14 : Analyse de la structure de l'arborescence, Fonds Simon) confirment les conclusions qui ont pu être tirées grâce aux visualisations d'*Archifiltre*. Les clés USB contiennent un grand nombre de *dossiers cylindriques* (ne comportant qu'un seul sous-dossier) avec 42 dossiers, soit un élément sur cinq. Pour la moitié d'entre eux, ils se trouvent à un niveau de répertoire élevé puisqu'il s'agit soit du dossier « content » (c'est-à-dire le premier dossier de l'arborescence, ce qui n'a rien d'étonnant), soit du dossier « FAT32 » (*File Allocation Table*, 32-bits), au second niveau de l'arborescence, correspondant au système de fichiers embarqué sur la plupart des supports amovibles (disquettes, clés USB, etc.) en raison de sa grande compatibilité entre les systèmes d'exploitation (*Windows*, *Mac*, *Linux*). Quant à la seconde moitié des dossiers cylindriques, il s'agit d'éléments système (« Store-V2 » et « Stores »), dans les derniers niveaux de l'arborescence (5-6).

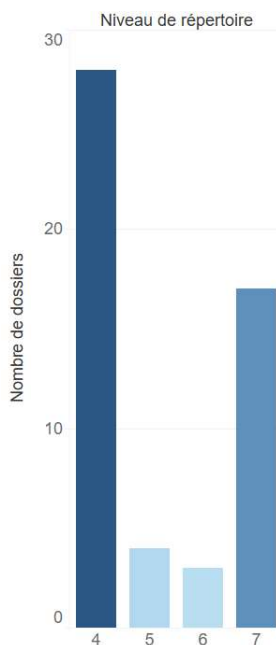
Les clés USB contiennent également beaucoup de *dossiers vides* (58 dossiers sur 189, soit 30 % du nombre total de dossiers), constitués exclusivement d'éléments liés au système de fichiers (il s'agit de dossiers intitulés « journals[...] » et stockés dans des répertoires « Spotlight-V100 » propres à Macintosh) se trouvant au septième et dernier niveau de l'arborescence. Une autre indication qui ressort des analyses statistiques permises grâce à l'extraction des métadonnées, c'est le faible nombre de *dossiers finaux non vides* puisqu'un quart seulement des dossiers contiennent *aucun* sous-dossier et *au moins* un fichier. Cela confirme donc également l'existence d'une arborescence très horizontale, comportant peu de subdivisions, et un nombre restreint de dossiers finaux qui contiennent alors la majorité des fichiers.

Tableau 14 : Analyse de la structure de l'arborescence, Fonds Simon

Structure de l'arborescence	Points d'analyse	Équation de recherche		Ana Simon	
		Dossier	Fichier	Nombre de dossiers	Proportion des dossiers totaux (%)
	Dossiers contenant un seul sous-dossier (cylindre)	1	0	42	22,22
	Dossiers vides (ni sous-dossier, ni fichier)	0	0	58	30,69
	Dossiers contenant des fichiers à la racine	$\geq 1$	$\geq 1$	35	18,52
	Dossiers finaux contenant un seul fichier	0	1	14	7,41
	Dossiers finaux avec au moins deux fichiers	0	$\geq 2$	38	20,11
	Dossiers finaux (total, avec dossiers vides)	0	$\geq 0$	110	58,20
	Dossiers finaux (total, sans dossiers vides)	0	$\geq 1$	52	27,51
	Dossiers structurés (non-finaux, sans fichiers racines, ni cylindres)	$\geq 2$	0	2	1,06
	Dossiers d'arborescence (total, avec cylindres)	$\geq 1$	$\geq 0$	79	41,80
	Total			189	100

Si l'on croise ces informations sur la structure de l'arborescence avec celles concernant sa profondeur, on peut facilement déterminer quel est le niveau à examiner en priorité. Dans le cas du Fonds Simon, c'est bien le quatrième niveau qui comporte le plus grand nombre de *dossiers finaux non-vides*, loin devant le septième et dernier niveau :

Figure 35 : Nombre de dossiers finaux non vides par niveau, Fonds Simon



#### 4.2.2.2 Francis Reusser

Tableau 15 : Analyse de la structure de l'arborescence, Fonds Reusser

Structure de l'arborescence	Points d'analyse	Équation de recherche		Francis Reusser	
		Dossier	Fichier	Nombre de dossiers	Proportion des dossiers totaux (%)
	Dossiers contenant un seul sous-dossier (cylindre)	1	0	606	15,71
	Dossiers vides (ni sous-dossier, ni fichier)	0	0	738	19,13
	Dossiers contenant des fichiers à la racine	$\geq 1$	$\geq 1$	588	15,24
	Dossiers finaux contenant un seul fichier	0	1	533	13,82
	Dossiers finaux avec au moins deux fichiers	0	$\geq 2$	1266	32,81
	Dossiers finaux (total, avec dossiers vides)	0	$\geq 0$	2537	65,76
	Dossiers finaux (total, sans dossiers vides)	0	$\geq 1$	1799	46,63
	Dossiers structurés (non-finaux, sans fichiers racines, ni cylindres)	$\geq 2$	0	126	3,27
	Dossiers d'arborescence (total, avec cylindres)	$\geq 1$	$\geq 0$	1320	34,21
	Total			3858	100

Le Fonds Reusser contient quelque 600 dossiers cylindriques, ce qui paraît élevé de prime abord. En y regardant de plus près cependant, on constate qu'il s'agit quasi exclusivement (485 sur 606) de dossiers se trouvant entre les niveaux 11 et 20 et qu'ils sont relatifs au matériel informatique de Francis Reusser (515 dossiers se partagent d'ailleurs 20 intitulés seulement). La liste des seuls intitulés significatifs (hors éléments « système ») est la suivante : « 24-28 MAI », « 22.mai », « 23.mai », « 25.mai », « 29.mai », « 30.mai », « DCIM », « IMAGES », « PHOTOS MANU-ZARETTI », « PHOTOS\_15\_18\_MAI », « PHOTOS\_MANU », « PHOTO-TOURNAGE-NH », « PRESSE », « RICOH GRANDE SALLE », « SANDOZ-FAMILY », « TOURNAGE SCOP ». Les cylindres sont donc composés de dossiers situés entre les niveaux 6 et 10 et sont relatifs aux photographies du tournage d'un film, au nom du photographe, ou à la date de la prise de vue (ainsi que du dossier « DCIM », *Digital Camera Images*, provenant du système de fichiers de l'appareil photographique). En outre, il est fréquent qu'un dossier cylindrique de niveau  $n$  soit suivi, hiérarchiquement parlant, d'un autre dossier cylindrique de niveau  $n-1$ , c'est-à-dire que les cylindres se développent le plus souvent sur plusieurs niveaux de répertoire, comme on peut le voir dans l'exemple ci-dessous. Une fois identifié, le traitement de ce type de cylindres (relativement peu fréquent de manière délibérée chez Francis Reusser) pourrait donc facilement être effectué (en supprimant les niveaux intermédiaires par exemple, afin de faire « remonter » l'arborescence de quelques niveaux et / ou en juxtaposant – voire en renommant – les intitulés des dossiers supérieurs afin qu'ils reflètent le contenu qui leur est subordonné).

Figure 36 : Dossiers cylindriques, Fonds Reusser,  
hfsplus\aaCINEATELIER\aaROUSSEAU\ROUSSEAU\_2012\IMAGES\PHOTOS\_NH



Le Fonds Reusser invite aussi à poser plus précisément la question des fichiers à la racine, dont il a été question précédemment – c’est-à-dire des fichiers qui ne se trouvent pas dans un dossier en bout d’arborescence, mais qui jouxtent des dossiers de même niveau. La présence de fichiers « à la racine » indique qu’un classement a bel et bien été entrepris mais que ce dernier n’est pas systématique, ni suffisamment précis puisque des fichiers sont subordonnés à un dossier général(-iste) qui a donné lieu, simultanément ou par la suite, à des sous-dossiers plus spécifiques dans lesquels les fichiers « à la racine » ne sont ou n’ont pas été intégrés (volontairement ou non). Le tableau ci-dessous donne un aperçu quantitatif des fichiers à la racine dans le Fonds Reusser jusqu’au niveau 13. En raison du lien de subordination qui existe entre un dossier (de niveau  $n$ ) qui contient des sous-dossiers et des fichiers à la racine au niveau inférieur ( $n-1$ ), nous avons privilégié la réalisation d’un tableau en « escalier » qui peut se lire de la manière suivante : en prenant les dossiers de niveau 3, on remarque que 58 dossiers sur les 108 que compte cet étage contiennent 599 fichiers à la racine au niveau inférieur ( $n-1$ , c’est-à-dire au niveau 4, sur les 1’004 fichiers que compte cet étage).

Tableau 16 : Statistiques relatives au nombre de fichiers à la racine, Fonds Reusser

Francis Reusser							
Profondeur		Dossiers			Fichiers		
Dossiers	Fichiers	Total	Avec des fichiers subordonnés à la racine	Proportion avec des fichiers subordonnés à la racine	Total	À la racine	Proportion de fichiers à la racine
Niveau 1		1	1	100,00			
	Niveau 2				14	15	107,14
Niveau 2		17	10	58,82			
	Niveau 3				89	70	78,65
Niveau 3		108	58	53,70			
	Niveau 4				1004	599	59,66
Niveau 4		382	134	35,08			
	Niveau 5				5505	2649	48,12
Niveau 5		607	177	29,16			
	Niveau 6				6479	1919	29,62
Niveau 6		1060	79	7,45			
	Niveau 7				6214	1378	22,18
Niveau 7		533	34	6,38			
	Niveau 8				3726	232	6,23
Niveau 8		174	53	30,46			
	Niveau 9				6645	283	4,26
Niveau 9		252	5	1,98			
	Niveau 10				4149	827	19,93
Niveau 10		67	17	25,37			
	Niveau 11				845	47	5,56
Niveau 11		151	15	9,93			
	Niveau 12				2516	65	2,58
Niveau 12		100	5	5,00			
	Niveau 13				60	10	16,67
Total		3452	588	17,03	37246	8094	21,73

Le tableau indique que le nombre absolu de dossiers contenant des fichiers à la racine au niveau directement inférieur est particulièrement élevé aux profondeurs 3 à 8, avec un pic aux niveaux 4 et 5 (de manière relative cependant, ce sont les tout premiers niveaux qui comptent proportionnellement le plus de dossiers contenant des fichiers à la racine au niveau subordonné). Pour ce qui est des fichiers, on constate que la majorité des fichiers à la racine se trouvent aux niveaux 5, 6 et 7 (avec respectivement 2'649, 1'919 et 1'378 fichiers) – proportionnellement en revanche ce sont les cinq premiers niveaux qui comptent le plus de fichiers à la racine vis-à-vis du nombre total de fichiers. Deux réserves, qui mettent en perspective ces chiffres (apparemment très) conséquents, s'imposent. Premièrement, la proportion de dossiers possédant des fichiers subordonnés à la racine (17,03 %), de même que la proportion de fichiers à la racine dans l'ensemble du fonds (21,73 %) ne sont pas si élevées. Deuxièmement, ces statistiques mériteraient d'être plus détaillées en fonction du niveau de répertoire et des dossiers individuels. À titre d'exemple, si l'on s'intéresse aux dossiers de niveau 4 qui comportent des fichiers subordonnés à la racine (134 sur 382, pour un total de 2'649 fichiers sur les 5'505 que compte le niveau 5), et que l'on se penche sur la distribution des fichiers au sein des dossiers, on obtient les résultats suivants : chaque dossier (n=4) contient *en moyenne* environ 20 fichiers à la racine (n=5), mais *la moitié* des dossiers (50%) contiennent 11 ou moins de fichiers subordonnés à la racine (et 33 dossiers ne contiennent même qu'un seul fichier à la racine, souvent composé d'un élément système « .DS\_Store »). La manière dont les fichiers à la racine sont globalement distribués dans les dossiers peut être résumée dans un tableau donnant entre autres la moyenne, la médiane et

les quartiles. On constate alors, de manière très générale, qu'il existe une grande majorité de dossiers contenant peu de fichiers à la racine (sur 588 dossiers, 156 n'en contiennent même qu'un seul !), et que quelques dossiers isolés contiennent beaucoup d'items (les répertoires de niveaux 4 à 6, c'est-à-dire les fichiers à la racine de niveau 5 à 7, mériteraient d'être analysés dans le détail au vu des résultats ci-dessous).

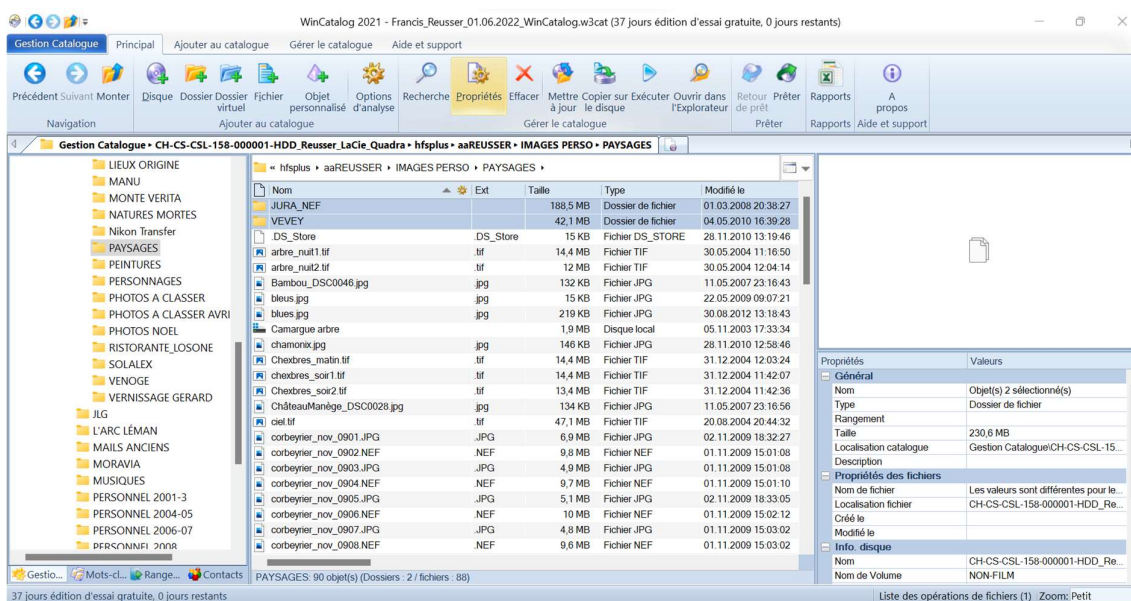
Tableau 17 : Distribution des fichiers à la racine par dossier, Fonds Reusser

Profondeur	Dossiers avec fichiers subordonnés	Nombre de fichiers subordonnés	Minimum	1er quartile	Médiane	Moyenne	3ème quartile	Maximum
Niveau 2	10	70	1,00	1,25	2,00	7,00	2,75	50,00
Niveau 3	58	599	1,00	1,00	1,00	10,33	5,75	167,00
Niveau 4	134	2649	1,00	2,00	11,00	19,77	31,00	146,00
Niveau 5	177	1919	1,00	3,00	6,00	10,84	13,00	88,00
Niveau 6	79	1378	1,00	2,50	8,00	17,44	31,00	75,00
Niveau 7	34	232	1,00	1,00	2,00	6,82	4,00	108,00
Niveau 8	53	283	1,00	2,00	6,00	5,34	6,00	26,00
Total	588	8904	1	1	5	13,77	15	428

Enfin, dernière indication, on peut regarder quel est le rapport entre la quantité de fichiers à la racine et le nombre de dossiers sur le même niveau de répertoire (on divise alors le nombre de fichiers à la racine par le nombre de dossiers de même niveau : si la valeur  $< 1$ , il y a plus de dossiers ; si la valeur  $= 1$ , il y a autant de fichiers que de dossiers ; si la valeur  $> 1$ , il y a plus de fichiers). Pour le Fonds Reusser, on obtient les proportions suivantes : dans un quart des cas (pour 156 dossiers, ce qui équivaut à 26,53 %) il y a plus de dossiers subordonnés que de fichiers à la racine ; dans environ 10 % des cas, le nombre de dossiers et de fichiers est équivalent (58 dossiers, soit 9,86 %) ; enfin, les deux-tiers environ des dossiers (374 dossiers, soit 63,60 %) contiennent plus de fichiers que de dossiers subordonnés.

En termes de traitement, les cas les plus extrêmes sont les plus facilement abordables : l'existence d'un seul fichier à la racine indique souvent la présence d'un élément lié au système informatique (de type « .DS\_Store ») ; à l'inverse, lorsqu'il y a (très) peu de dossiers et énormément de fichiers, il peut s'agir notamment d'erreurs (comme on pouvait le voir sur la Figure 36, avec la présence de 428 images et de 2 dossiers vides intitulés « dossier sans titre » et « dossier sans titre 2 »), ou de tris sélectifs / de rassemblements thématiques opérés par Francis Reusser lui-même (on le voit par exemple sur la figure ci-dessous).

Figure 37 : Fichiers à la racine et regroupements thématiques, *WinCatalog*, Fonds Reusser

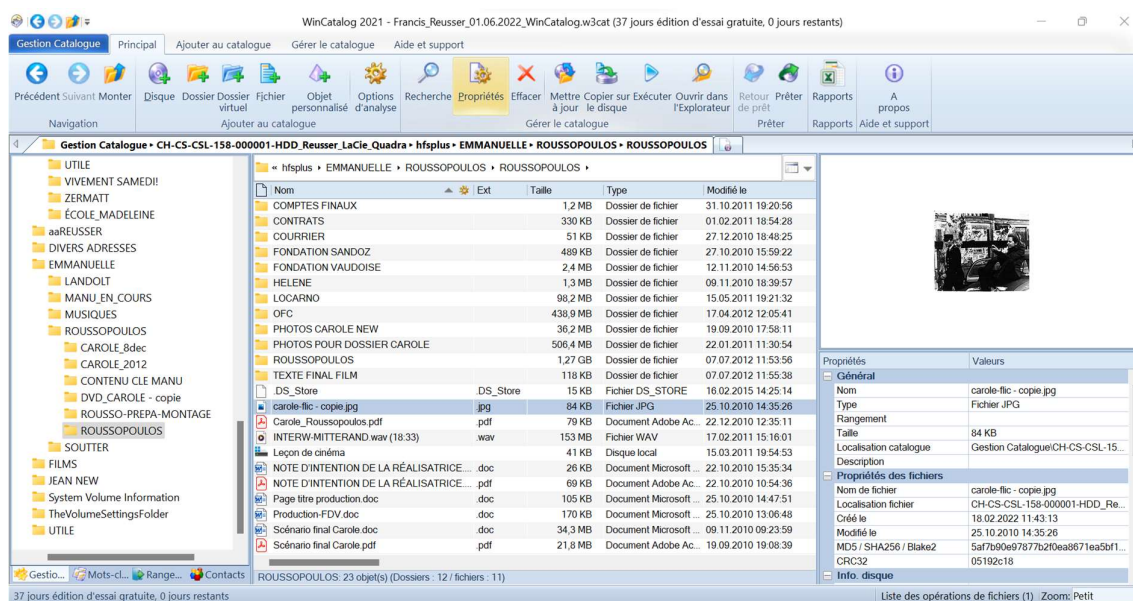


Grâce aux quelques pointages réalisés, il semblerait que ce cas de figure (regroupement thématique ou sélection) se rencontre surtout aux niveaux intermédiaires de l'arborescence (4 à 6), lorsque le nombre de fichiers est très élevé, et que cela concerne principalement les répertoires de photographies.

Les cas les plus difficiles en termes d'évaluation et de classification, ce sont les dossiers comportant plusieurs sous-dossiers subordonnés *et* plusieurs fichiers à la racine. Le dossier ci-dessous est un bon exemple puisqu'il y a 12 dossiers et 11 fichiers, et que ces derniers sont de formats divers (« .pdf », « .doc », « .wav ») et de natures différentes (on a des textes, des images et des fichiers audio). Ce dernier cas de figure n'est heureusement pas le plus fréquent dans le Fonds Francis Reusser, puisqu'il y a relativement peu de dossiers comportant des fichiers significatifs à la racine, ce qui témoigne d'une volonté d'organisation interne et de l'existence d'un classement qui a été entrepris plutôt minutieusement par le créateur des documents.



Figure 38 : Fichiers à racine et dossiers, exemple de co-présence, *WinCatalog*, Fonds Reusser



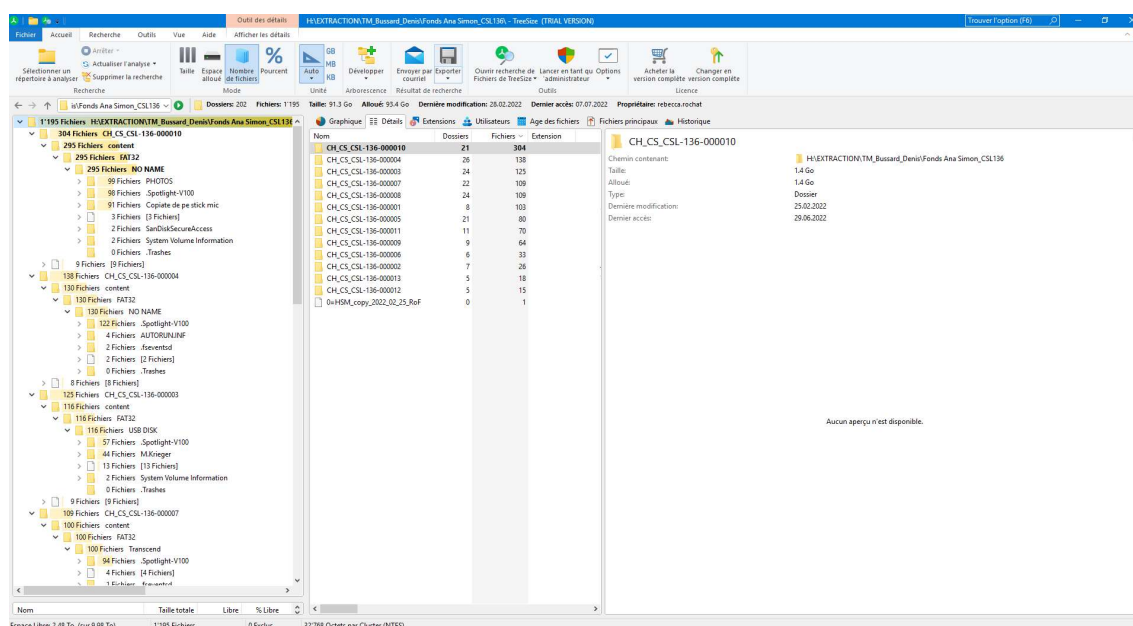
### 4.2.3 Analyse du « plan de classification » initial

Grâce aux enseignements tirés jusque-là, et bénéficiant de la connaissance de la structure générale de l'arborescence, on est en mesure de s'intéresser plus précisément à la manière dont les éléments sont organisés et comment ils font apparaître le plan de classification initial du créateur de documents. Cette analyse peut être menée en étudiant la nomenclature des dossiers, via le développement sélectif de l'arborescence par niveau de profondeur ou en étudiant les intitulés de dossiers grâce à l'extraction des métadonnées. On peut alors automatiquement extraire la liste complète des noms de dossiers (« Intitulés de tous les dossiers ») et la traiter pour faire apparaître les intitulés singuliers (« Intitulés uniques »). Cela permet d'étudier un nombre restreint de dossiers et d'analyser plus finement leurs intitulés, de manière à dégager les caractéristiques propres au répertoire. Nous donnons ci-dessous un exemple d'analyse avec le Fonds Ana Simon.

Sachant que les trois premiers niveaux de chaque clé USB comportent un seul dossier à l'intitulé identique (« content », « FAT32 », « NO NAME »), nous étudierons en priorité les dossiers du quatrième échelon. Pour cela, l'on peut utiliser la fonctionnalité proposée par *TreeSize* qui consiste à développer l'arborescence par niveau de profondeur. Il est ainsi possible de déployer l'arborescence de sorte à ce que tous les dossiers d'une même profondeur apparaissent simultanément. L'avantage de cette fonctionnalité, c'est qu'elle permet de consulter les dossiers d'un même niveau dans leur contexte, c'est-à-dire de visualiser leurs dossiers parents *et* leurs voisins directs, mais aussi de comparer un même niveau dans plusieurs branches du répertoire. Dans le Fonds Ana Simon par exemple, on peut ainsi comparer tous les dossiers de quatrième niveau au sein d'une clé USB *et* dans les treize clés USB dans une seule fenêtre.



Figure 39 : Développement sélectif de l'arborescence, niveau 4, *TreeSize*, Fonds Simon



Si l'on désire poursuivre l'analyse, on peut étudier l'intitulé des dossiers via l'extraction des métadonnées. Vingt-deux intitulés différents sont utilisés pour l'identification des 49 dossiers qui composent le niveau 4 du Fonds Simon. Six noms sont partagés par 33 dossiers tandis que 16 dossiers possèdent des noms uniques. Parmi ces 22 intitulés différents, 14 nous paraissent significatifs, c'est-à-dire qu'ils renseignent sur le contenu du dossier et peuvent avoir été nommés ainsi par la productrice des documents. Quatre catégories peuvent être distinguées : des intitulés renvoyant à la typologie des documents (comme « PHOTOS », « VIDEOS\_TS »), à des personnalités actives dans le domaine culturel (« benjamin fondane », « Jeanne\_moreau »), à des projets ou à des thèmes (dont « la petite vendeuse de lampe » ou « sortilege de genève », films réalisés par Ana Simon) ainsi que des noms livrant des indices de « méta-classement » nous renseignant sur l'utilisation des documents ou sur leur finalité (« ana a imprimer » ou « Copiate de pe stick mic »).

Une étude similaire pour les dossiers de niveau 5 livre les résultats suivants : 21 dossiers, avec 7 intitulés uniques et 3 dossiers significatifs (« Alte Foto », « ANA SIMON 2016 » et « Rezerva Ap Foto », dans les clés 3 et 10) tandis que le niveau 6 est construit avec 16 dossiers, dont 11 intitulés uniques et 3 dossiers significatifs (« ana a imprimer », « photos ana espagne 2015 » et « SOLEIL », tous dans la clé 10). On constate alors que les niveaux 5 et 6 apportent peu d'informations nouvelles : les catégories sont les mêmes (typologie, méta-classement, projet-thèmes) et ne contribuent donc pas à une classification plus fine ; et les intitulés significatifs du niveau 6 sont identiques à certains noms que l'on trouve au niveau 4. Enfin, si l'on considère le fonds entier, on obtient les valeurs suivantes : 189 dossiers en tout, avec 66 intitulés uniques, et seulement 21 dossiers qui possèdent un intitulé significatif (1 au 3<sup>e</sup> niveau ; 14 au 4<sup>e</sup> niveau ; 3 au 5<sup>e</sup> et au 6<sup>e</sup> ; 0 au 7<sup>e</sup> niveau).

En conclusion, le Fonds Ana Simon est constitué de quatre grandes catégories le plus souvent juxtaposées au quatrième niveau (il n'y a donc pas de classification primaire, secondaire, etc.). À ce stade de l'analyse, il nous paraîtrait recommandé de commencer l'évaluation par les catégories les moins « renseignantes », c'est-à-dire les dossiers de « méta-classement » et

de typologie (les dossiers par projet et thème sont les plus susceptibles d'être conservé en l'état).

Tableau 18 : Nomenclature des dossiers, niveau 4, Fonds Simon

Fonds Ana Simon, dossiers de niveau 4					
Clé	Intitulés de tous les dossiers	Intitulés uniques	Occurrences	Dossiers significatifs	Clé
13	49	22	49	14	8
1	ana a imprimer	ana a imprimer	1	ana a imprimer	1
1	photos ana espagne 2015	AUTORUN.INF	1	photos ana espagne 2015	1
1	SanDiskSecureAccessV2.0	benjamin fondane	1	SOLEIL	1
1	SOLEIL	Copiate de pe stick mic	1	Divers	2
1	System Volume Information	Divers	1	Prod Ana	2
2	.Trashes	FOUND.000	1	M.Krieger	3
2	Divers	Jeanne moreau	1	la petite vendeuse de lampe	6
2	FOUND.000	la petite vendeuse de lampe	1	sortilege de genève	6
2	Prod Ana	LES MUSES ENDORM	1	VIDEO_TS	6, 13
3	.Spotlight-V100	M.Krieger	1	Jeanne moreau	8
3	.Trashes	PHOTOS	1	Copiate de pe stick mic	10
3	M.Krieger	photos ana espagne 2015	1	PHOTOS	10
3	System Volume Information	Prod Ana	1	benjamin fondane	11
4	.fseventsd	SanDiskSecureAccess	1	LES MUSES ENDORM	11
4	.Spotlight-V100	SOLEIL	1		
4	.Trashes	sortilege de genève	1		
4	AUTORUN.INF	VIDEO_TS	2		
5	.fseventsd	SanDiskSecureAccessV2.0	3		
5	.Spotlight-V100	System Volume Information	4		
5	.Trashes	.fseventsd	6		
5	System Volume Information	.Spotlight-V100	8		
6	la petite vendeuse de lampe	.Trashes	10		
6	sortilege de genève				
6	VIDEO_TS				
7	.fseventsd				
7	.Spotlight-V100				
7	.Trashes				
8	.fseventsd				
8	.Spotlight-V100				
8	.Trashes				
8	Jeanne moreau				
9	.fseventsd				
9	.Spotlight-V100				
9	.Trashes				
10	.Spotlight-V100				
10	.Trashes				
10	Copiate de pe stick mic				
10	PHOTOS				
10	SanDiskSecureAccess				
10	System Volume Information				
11	.Spotlight-V100				
11	.Trashes				
11	benjamin fondane				
11	LES MUSES ENDORM				
11	SanDiskSecureAccessV2.0				
12	.fseventsd				
12	.Trashes				
13	SanDiskSecureAccessV2.0				
13	VIDEO_TS				

**Légende des couleurs**

**Méta-classement**

**Typologie**

**Projet-thème**

**Personnalité**

Tableau 19 : Arborescence et volumétrie, tâches

Tâche	Sous-tâches	Méthodes et instruments	Points d'analyse	Outils
Arborescence et volumétrie	Analyse de la profondeur de l'arborescence	<ul style="list-style-type: none"> <li>* Visualisation</li> <li>* Navigation sélective</li> <li>* Recherche avancée</li> <li>* Récolement (analyse statistique)</li> </ul>	<ul style="list-style-type: none"> <li>* Profondeur maximale (nombre total de niveaux)</li> <li>* Nombre de dossiers par niveau (forme générale de l'arborescence)</li> <li>* Profondeur des dossiers finaux</li> <li>* Profondeur par répertoire(s)</li> <li>* Homogénéité - hétérogénéité de la profondeur</li> </ul>	<ul style="list-style-type: none"> <li>* <i>Archifiltre</i> (visualisation)</li> <li>* <i>TreeSize Professional</i> (développement sélectif et recherche avancée)</li> <li>* Karen's Directory Printer (récolement)</li> <li>* <i>Excel</i> (analyse statistique)</li> <li>* <i>Tableau</i> (visualisations graphiques)</li> </ul>
	Analyse de la structure de l'arborescence		<ul style="list-style-type: none"> <li>* Dossiers contenant un seul sous-dossier (cylindres)</li> <li>* Dossiers vides (ni sous-dossier, ni fichier)</li> <li>* Dossiers contenant des fichiers à la racine</li> <li>* Dossiers finaux contenant un seul fichier</li> <li>* Dossiers finaux avec au moins deux fichiers</li> <li>* Dossiers finaux (total, avec dossiers vides)</li> <li>* Dossiers finaux (total, sans dossiers vides)</li> <li>* Dossiers structurés (non-finaux, sans fichiers racines, ni cylindres)</li> <li>* Dossiers d'arborescence (total, avec cylindres)</li> <li>* Nombre de fichiers à la racine (par profondeur)</li> <li>* Nombre moyen de fichiers par dossier</li> </ul>	
	Analyse du plan de classification initial		<ul style="list-style-type: none"> <li>* Mode de classement primaire</li> <li>* Mode de classement secondaire</li> <li>* Nomenclature des éléments (système de nommage)</li> </ul>	

## 4.3 Traitement des dossiers vides

Les dossiers vides sont des répertoires ne contenant ni sous-dossier ni fichier subordonné. S'ils peuvent s'avérer relativement nombreux, à cause des dossiers créés automatiquement par certains systèmes de fichiers ou par des applications installées sur nos ordinateurs, il convient toutefois de faire preuve de prudence avant de procéder à leur suppression massive. Une étape d'identification suivie par une phase d'analyse sont nécessaires avant de décider du sort final de ces dossiers spécifiques. On accordera notamment une attention particulière aux caractéristiques suivantes : les dossiers possèdent-ils des intitulés significatifs, qui ont été formulés par le créateur des documents ou qui renseignent sur ses activités ? Existe-t-il au sein du répertoire analysé d'autres dossiers (non vides) possédant le même intitulé ? Quel est l'emplacement, au sein de l'arborescence complète, du dossier vide, et quels sont ses « voisins » directs (s'agit-il également de dossiers vides ou y a-t-il des dossiers possédant la même typologie d'intitulé – par thème, par date, par activité, etc. ?). Tous ces points d'analyse permettront de prendre une décision éclairée : doit-on conserver le dossier en l'état ou peut-on légitimement supprimer l'élément vide sans perte d'information importante (et dans quelle mesure faut-il documenter sa suppression) ?

Tableau 20 : Traitement des dossiers vides, tâches

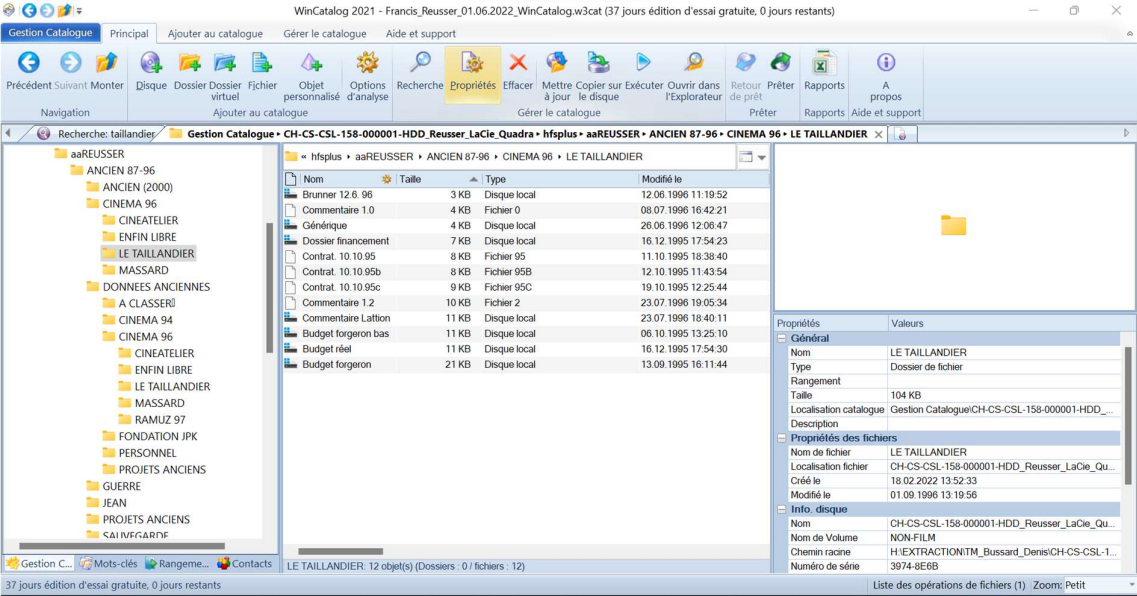
Tâche	Sous-tâches	Méthodes et instruments	Points d'analyse	Outils
Traitement des dossiers vides	Identification des dossiers vides	* Recherche avancée * Récolement	* Profondeur de la recherche (avec / sans fichiers système) * Quantité de dossiers vides * Intitulés identiques	* TreeSize Professional (recherche avancée-générale, suppression) * Karen's Directory Printer (récolement)
	Analyse individuelle et décision du sort final	* Recherche générale * Analyse de l'arborescence	* Nomenclature (significative et porteuse d'informations) * Unicité de l'intitulé / Existence d'un dossier non vide sur le même sujet * Emplacement dans l'arborescence (chemin parent et voisinage direct)	
	Application du sort final	* Récolement	* Conservation justifiée / injustifiée * Suppression avec documentation * Suppression sans documentation	

À titre d'exemple, le Fonds Reusser compte un peu plus de 700 dossiers vides (720 via la recherche avancée de *TreeSize*, 738 selon le récolement de *Karen's Directory Printer*), qui se partagent presque 200 intitulés distincts<sup>17</sup>. Hormis les nombreux dossiers système, quelques intitulés attirent l'attention dont « L'AUTRE », « LE TAILLANDIER » et « MONTAGNE ». Ces trois dossiers possèdent un titre en majuscules (or la casse est significative chez Reusser) et se trouvent tous dans le même répertoire principal : « aaREUSSER/ANCIEN 87-96 », puis dans les embranchements « DONNEES ANCIENNES\PROJETS ANCIENS », « DONNEES ANCIENNES\CINEMA 96 » et « SAUVEGARDE SE\DONNEES ». En effectuant une recherche dans tout le fonds sur la base de ces titres-là, on constate qu'il existe d'autres dossiers avec les mêmes intitulés, conservés entre autres dans le même répertoire principal. Un dossier intitulé « LE TAILLANDIER » se trouve effectivement à deux reprises dans un double répertoire « CINEMA 96 » ; mais si l'un est vide, l'autre contient en revanche des documents relatifs au film *Le Taillandier* : *Charles Maire, forgeron*, tourné par Francis Reusser et Emmanuelle de Riedmatten en 1996 à Troistorrents (voir Figure 40). Le même cas de figure se présente pour les dossiers intitulés « L'AUTRE » et « MONTAGNE », avec des dossiers vides et des dossiers contenant des documents (dont un synopsis pour « L'AUTRE »). L'existence d'autres dossiers à la nomenclature identique et porteurs de documents, ainsi

<sup>17</sup> Ces chiffres concernent les dossiers « totalement » vides et ne prennent donc pas en compte les dossiers qui ne contiennent par exemple qu'un seul fichier système – le chiffre serait donc probablement plus conséquent si la recherche de dossiers vides était effectuée après l'élimination des éléments système du répertoire.

qu'un contexte d'enregistrement très similaire (même répertoire principal et dossiers parents de même typologie) permettent la suppression des dossiers vides, sans qu'il soit nécessaire de documenter l'élimination. Ces exemples illustrent le type de questions que l'on doit se poser avant une élimination, en isolant les titres significatifs et porteurs d'informations, en étudiant leur contexte direct (ici, le répertoire « PROJETS ANCIENS » aurait pu justifier la conservation du dossier s'il avait été question d'un projet cinématographique avorté, dont c'est la seule trace conservée...), et en menant une brève enquête dans le reste du fonds pour décider de leur sort final.

Figure 40 : Dossiers vide et non-vide, « LE TAILLANDIER », WinCatalog, Fonds Reusser



## 4.4 Traitement des redondances strictes

Le traitement des redondances strictes fait partie des tâches prioritaires lorsqu'il s'agit d'évaluer un fonds d'archives numériques, et le « dédoublonnage » est fréquemment mentionné dans les flux de travail existants. La multiplication d'éléments identiques au sein d'un fonds d'archives est effectivement facilitée par la dématérialisation des contenus : alors qu'il est nécessaire de photocopier les documents analogiques, d'en recevoir plusieurs copies par des tiers, ou d'en acquérir différents exemplaires – autant d'actes qui requièrent un investissement temporel ou financier, une place de stockage considérable, ainsi qu'une sélection en amont des contenus à dupliquer –, les documents numériques, de par leur dématérialisation, peuvent être générés ou copiés (et diffusés) en un nombre infini d'exemplaires sans que cela demande un investissement conséquent. L'hypothèse pourrait même être renversée : il peut sembler parfois plus chronophage de retrouver un document au sein d'une arborescence complexe et ramifiée, que d'en produire une copie dans l'espace de travail en cours d'utilisation.

### 4.4.1 Traitement des redondances : aperçu général

Lorsqu'on évalue un fonds d'archives numériques, il convient en premier lieu de s'interroger sur le nombre total de redondances que l'on peut trouver au sein du répertoire analysé. La quantité d'éléments identiques est un premier indicateur de la pratique du créateur des documents : un nombre important de redondances peut être le signe d'une pratique récurrente, voire systématique. Les redondances ne seraient alors pas fortuites, accidentelles (fausse manipulation) ou techniques (générées informatiquement par exemple), mais témoigneraient d'une manière de travailler et d'utiliser le support de stockage. En effet, si l'arborescence est une projection de l'univers mental du créateur des documents, les redondances en sont une manifestation moins évidente (plus dissimulée), et peuvent révéler un système sous-jacent à l'organisation directement déchiffrable via la nomenclature des dossiers.

Dans le Fonds Reusser par exemple, les redondances sont fréquentes comme on peut le constater dans le tableau récapitulatif ci-dessous. Retenons en particulier les chiffres suivants : le fonds compte 3'948 dossiers en tout, dont 3'626 dossiers distincts (91,8 %) ; et il contient 38'706 fichiers en tout, dont 25'833 fichiers différents (66,7 %). Dans le détail, notons qu'il y a 144 dossiers qui possèdent au moins un exemplaire redondant (3,6 % de tous les dossiers possèdent donc au moins un exemplaire identique dans le fonds) et qu'il y a 6'051 fichiers dont il existe une ou plusieurs copies (15,6 %).

Tableau 21 : Recherche de redondances strictes : cas pratiques, résumé

Variables	Fonds Francis Reusser				Fonds Ana Simon			
	AllDup	Archifiltre	TSP	WinCatalog	AllDup	Archifiltre	TSP	WinCatalog
Dossiers : nombre total de redondances	x	926	466	x	x	9	5	x
Dossiers : nombre de groupes/lots	x	x	144	x	x	x	2	x
Dossiers : volume total de redondances	x	x	58.7 Go	x	x	x	6.7 Mo	x
Dossiers : nombre de dossiers distincts	x	x	3626	x	x	x	199	x
Dossiers : nombre de redondances	x	x	322	x	x	x	3	x
Fichiers : nombre total de redondances	21662	18408	18924	24393	433	390	433	530
Fichiers : nombre de groupes/lots	6268	6002	6051	x	139	140	139	x
Fichiers : volume total de redondances	121,99 Go	82,6 Go	119,9 Go	x	189,96 Mo	110,1 Mo	190 Mo	x
Fichiers : nombre de fichiers distincts	x	25183	25833	x	x	940	900	x
Fichiers : nombre de redondances	15394	12406	12873	x	294	250	294	x

Les chiffres ci-dessus<sup>18</sup> proviennent très majoritairement des logiciels testés et en particulier de *TreeSize* qui fournit entre autres le nombre total de redondances (relativement peu pertinent puisqu'il additionne une instance de chaque élément présent en plusieurs exemplaires *et* le nombre total de ses duplicatas) et le nombre des duplicatas (soit le nombre total de redondances auquel on soustrait une occurrence de chaque élément) ; le nombre d'occurrences uniques qui possèdent au moins un double doit en revanche être calculé (c'est le nombre de « groupes », de « lots », que fournit notamment *Archifiltre* pour les fichiers).

L'affichage des résultats par onglet regroupant les instances d'un même élément livre des informations sur le nombre d'exemplaires de chaque dossier / fichier redondant, mais les logiciels testés ne fournissent malheureusement pas de statistiques à ce propos. De même, si l'on peut trier les groupes par « Chemin contenant », afin de faire apparaître facilement les groupes dont les instances se trouvent toutes dans le même répertoire, seule l'indication « [plusieurs] » apparaît pour les instances enregistrées dans des chemins parents distincts. Les logiciels ne proposent finalement qu'un traitement au niveau *micro*, c'est-à-dire au niveau de l'élément individuel et de ses duplicatas, et obligent à trancher les différents cas de figure individuellement. Cette approche est extrêmement chronophage et n'est guère envisageable lorsqu'on est confronté à un nombre important de redondances (l'on devrait par exemple traiter au cas par cas les quelque 6'000 lots de duplicatas qui se trouvent dans le Fonds Reusser). Si les logiciels offrent bel et bien des outils de « sélection » semi-automatisée des éléments (par taille, date, nom, chemin, etc.), encore faut-il savoir quel filtre appliquer... Comme nous le soulignons précédemment (voir 3.3.3.3 Recherche de redondances strictes : conclusion), il manque donc des instruments d'analyse *macro* pour comprendre de quelle manière les redondances sont « distribuées » dans le fonds d'archives et tenter d'identifier leur « organisation » interne. Les méthodologies, graphiques et instruments que nous proposons ci-dessous pour l'analyse des redondances peuvent s'appliquer aussi bien aux dossiers qu'aux fichiers. À ce stade de la recherche, nous avons privilégié l'étude des dossiers redondants du Fonds Reusser en raison de leur nombre plus restreint – avec des compétences limitées dans le traitement de données et dans les langages de programmation, la plupart des analyses ont été réalisées manuellement, mais elles pourraient faire l'objet d'une automatisation et figurer si nécessaire dans le cahier des charges d'un nouvel outil.

Pour ce qui est du flux de travail général (voir Tableau 27 : Traitement des redondances strictes, tâches), nous proposons de procéder du général au particulier, en commençant par analyser et traiter les *dossiers* redondants avant de s'atteler aux *fichiers* en plusieurs exemplaires, en débutant systématiquement par le traitement des redondances qui se trouvent dans le même répertoire (dossier parent / chemin contenant identique) puis en analysant les redondances dont les éléments sont enregistrés dans des répertoires différents (dossiers parents / chemins contenant distincts). Si tous les logiciels ne proposent pas la recherche de dossiers redondants, c'est pourtant une étape prioritaire, afin de mener une évaluation *top-down* et de dégrossir le nombre de fichiers à analyser (voir chapitre 3.3.3.2.2 *TreeSize*).

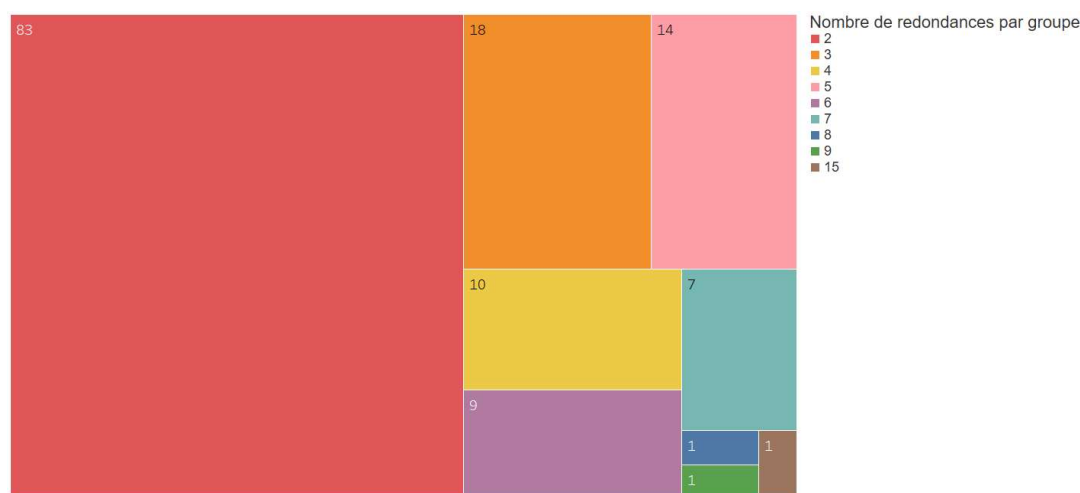
Afin de disposer d'une vision panoramique de la situation, il convient de s'interroger sur le nombre d'exemplaires que compte chaque groupe de redondances identifiées. Cela permet de compléter les statistiques générales sur le nombre total de redondances et le nombre d'éléments qui se trouvent en plusieurs exemplaires dans le répertoire. Dans le Fonds

---

<sup>18</sup> Les valeurs en rouge sont des champs calculés, c'est-à-dire des valeurs qu'il a été possible de calculer à partir des autres informations statistiques données par les logiciels.

Reusser, on constate que la majorité des dossiers redondants sont présents en deux exemplaires dans le fonds (83 groupes sur 144, soit environ 57 % – il s'agit du rectangle rouge dans le graphique ci-dessous). Les dossiers présents de (très) nombreuses fois ont plus de chances d'avoir été générés de manière involontaire, ou de faire partie de lots massivement copiés d'un répertoire à l'autre de façon systématique ; ils pourront donc faire l'objet d'une première évaluation, moins complexe, et qui aura l'avantage de traiter en une fois un grand nombre d'éléments. Dans le cas de Reusser, les dossiers les plus redondants concernent le matériel informatique (manuels d'utilisation, licences, codes d'accès, etc.) acquis par la maison de production du réalisateur (*Ciné-atelier*), et ces dossiers sont copiés d'une année sur l'autre dans les dossiers administratifs de la société. Quant aux quinze occurrences d'un même dossier (rectangle brun dans le graphique ci-dessous), il s'agit en réalité de répertoires aux intitulés très divers qui ne contiennent qu'un seul fichier système identique. On concentrera donc notre attention sur les dossiers présents en deux exemplaires seulement – ces derniers sont non seulement les plus nombreux, mais ils possèdent aussi les intitulés les plus significatifs et sont donc susceptibles d'avoir été générés volontairement par le créateur des documents, ce qui justifierait une analyse plus détaillée avant leur traitement massif.

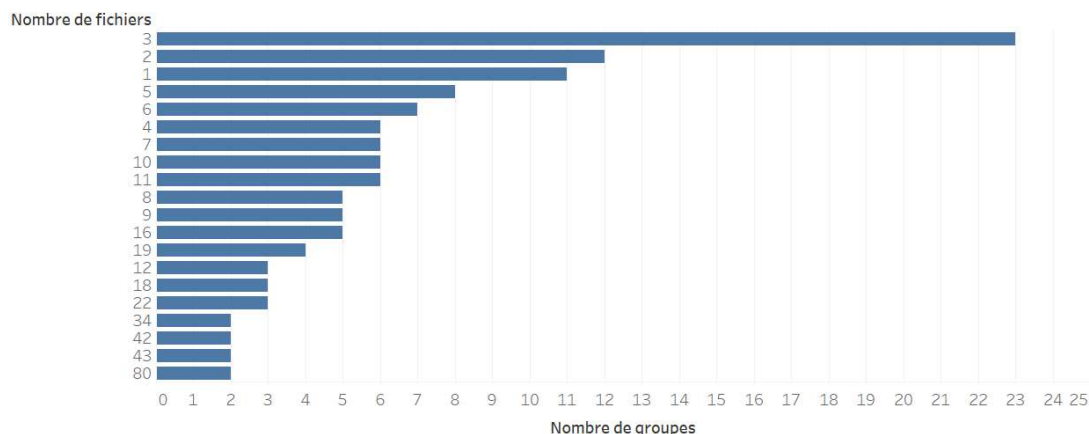
Figure 41 : Nombre d'occurrences de chaque dossier redondant, par groupe, Fonds Reusser



Chez Reusser, les dossiers en plusieurs exemplaires contiennent en tout 9'473 fichiers ; en ne conservant qu'un seul exemplaire de chaque dossier (144, un par groupe), on ne garderait « que » 3'221 fichiers, et on supprimerait 6'252 fichiers, qui n'apparaîtraient ainsi plus dans les résultats de recherche sur les fichiers redondants. Cette statistique générale ne nous dit rien en revanche de la composition exacte des dossiers redondants. Pourtant, il peut être intéressant de connaître la répartition des fichiers par groupe de redondances. Autrement dit : combien de fichiers comportent les dossiers dont il existe plusieurs exemplaires ? A-t-on affaire à des copies de dossiers très complets, contenant énormément d'éléments, ou s'agit-il de dossiers relativement minces ? Comme on peut le voir sur la figure ci-dessous, les dossiers redondants du Fonds Reusser contiennent souvent peu de fichiers subordonnés : 23 groupes de dossiers sur 144 contiennent seulement 3 fichiers et la moitié des dossiers redondants comptent jusqu'à 7 fichiers seulement. Les dossiers sont donc généralement peu garnis, à l'exception de quelques répertoires qui contiennent plusieurs centaines de fichiers (il s'agit alors généralement de photographies).



Figure 42 : Nombre de fichiers contenus dans un dossier redondant, Fonds Reusser

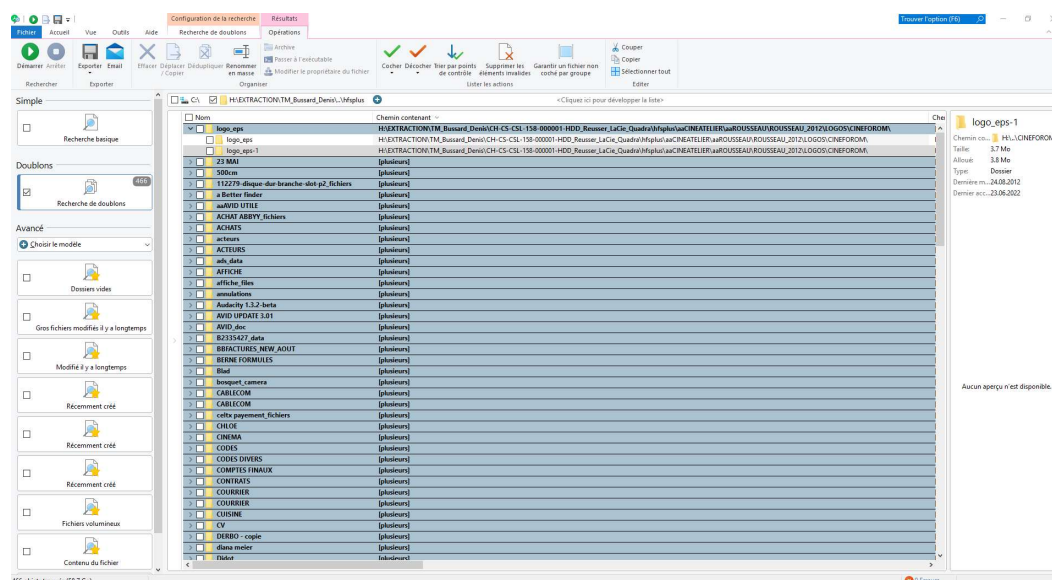


#### 4.4.2 Traitement des redondances : analyse des répertoires parents

Jusque-là, les analyses concernent les redondances dans leur ensemble, de sorte à en donner une première image (on sait dorénavant combien d'exemplaires il existe de dossiers redondants, et combien de fichiers ils comportent). Pouvoir les traiter nécessite en revanche des analyses qui prennent en compte non plus seulement les éléments individualisés, mais leur *contexte d'enregistrement*. Il s'agit de comprendre où se trouvent ces redondances, dans quel(s) répertoire(s) elles figurent, et quel(s) lien(s) elles entretiennent.

Deux grands cas de figure se présentent alors : les dossiers identiques se trouvent au sein du *même* répertoire ; les dossiers redondants appartiennent à des branches de l'arborescence différentes. Le premier cas est de loin le plus simple, mais il se présente relativement rarement. Dans le Fonds Reusser, pour les dossiers, on ne compte qu'un seul groupe de redondances dont les deux occurrences se trouvent *exactement* au même endroit. Un tri sur la colonne « Chemin contenant » dans *TreeSize* permet de les faire apparaître en tête des résultats de recherche.

Figure 43 : Dossiers redondants possédant le même répertoire parent, *TreeSize*, Fonds Reusser

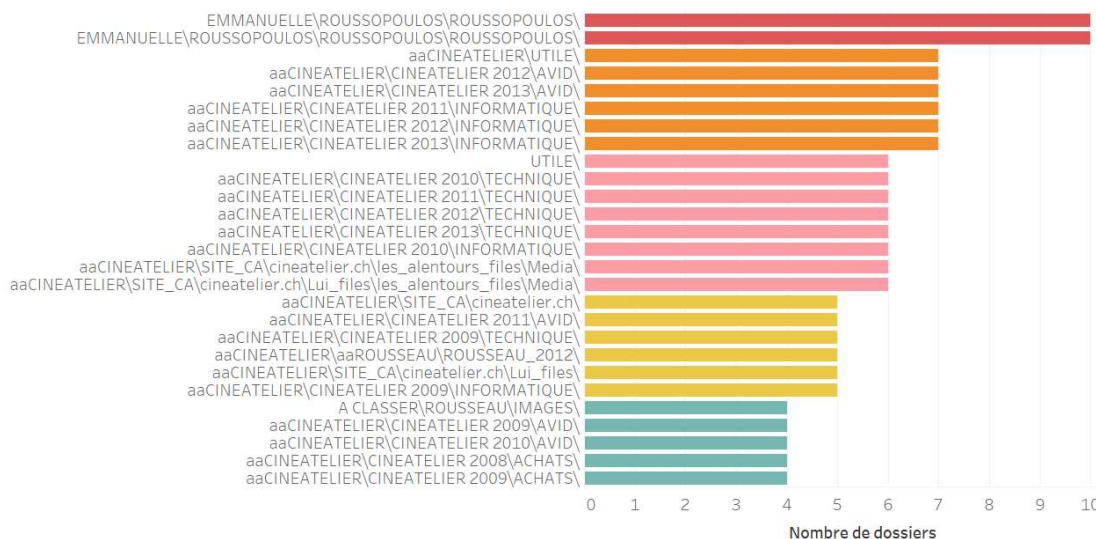


Dans ce cas-là, lorsque les dossiers redondants ont le même répertoire parent, la décision quant à leur sort final est facile à prendre : l'on peut sans perte d'informations ni de contexte supprimer l'une des deux occurrences. L'intitulé des dossiers ainsi que les dates de dernière modification (s'ils diffèrent) pourront être comparés afin de déterminer l'élément à préserver (on conservera l'intitulé le plus significatif ou porteur d'informations ainsi que le dossier le plus récent).

Le second cas de figure, lorsque les redondances se trouvent dans des branches différentes de l'arborescence, est autrement plus complexe, puisque cela nécessite une analyse fine des liens qui existent entre les éléments, afin de mettre au jour un (potentiel) système propre au créateur, et de déterminer le sort final à réserver aux éléments : doit-on conserver plusieurs occurrences ou peut-on n'en garder qu'une seule (et si oui, laquelle) ? Il est alors nécessaire d'étudier les *répertoires parents* (isolés dans la colonne « Chemin contenant ») des dossiers redondants.

Dans un premier temps, il peut être intéressant d'identifier, de manière générale, les répertoires qui accueillent le plus grand nombre de dossiers dont il existe plusieurs occurrences dans le fonds d'archives. Cette analyse ne prend donc pas (encore) en compte les liens qui existent entre deux ou plusieurs occurrences d'un même groupe de redondances ; il s'agit seulement d'identifier les répertoires *contenant* des dossiers redondants, quel que soit le(s) lieu(x) dans le(s)quel(s) se trouvent les autres occurrences du même dossier. Pour ce faire, on va comptabiliser le nombre de fois qu'un chemin parent apparaît dans la liste complète des dossiers redondants. L'illustration ci-dessous présente les résultats pour le Fonds Reusser (les répertoires sont classés par nombre de dossiers non-uniques hébergés, puis par niveau de profondeur dans l'arborescence – pour des questions de lisibilité, nous n'avons conservé ici que les répertoires comptant quatre dossiers au minimum).

Figure 44 : Nombre de dossiers redondants par répertoire parent, Fonds Reusser



Certaines branches de l'arborescence semblent donc contenir plusieurs dossiers non-uniques, comme c'est le cas du répertoire « EMMANUELLE\ROUSSOPOULOS\ROUSSOPOULOS\ ». Les dossiers redondants qu'il contient appartiennent à des « groupes » différents (comme on peut le voir dans la figure ci-dessous, avec des identifiants de groupes numérotés de 131 à 143), mais cela signifie qu'au moins un exemplaire de chaque groupe est stocké dans le

répertoire mentionné. Alors que le répertoire entier comprend 12 dossiers, cela signifie que plus de 80 % de son contenu (en termes de dossiers) existe ailleurs dans le Fonds Reusser.

Tableau 22 : Éléments non-uniqes au sein d'un répertoire, Fonds Reusser

ID de groupe	Nbr d'ex.	Nom	Chemin contenant	Fichiers	Taille	Profondeur
131	2	COMPTES FINAUX	EMMANUELLE\ROUSSOPOULOS\ROUSSOPOULOS\	3	1.2 Mo	4
132	2	CONTRATS	EMMANUELLE\ROUSSOPOULOS\ROUSSOPOULOS\	3	330.0 Ko	4
133	2	COURRIER	EMMANUELLE\ROUSSOPOULOS\ROUSSOPOULOS\	3	50.5 Ko	4
134	2	FONDATION SANDOZ	EMMANUELLE\ROUSSOPOULOS\ROUSSOPOULOS\	3	488.5 Ko	4
135	2	FONDATION VAUDOISE	EMMANUELLE\ROUSSOPOULOS\ROUSSOPOULOS\	12	2.4 Mo	4
136	2	HELENE	EMMANUELLE\ROUSSOPOULOS\ROUSSOPOULOS\	3	1.3 Mo	4
137	2	LOCARNO	EMMANUELLE\ROUSSOPOULOS\ROUSSOPOULOS\	7	98.2 Mo	4
141	2	PHOTOS CAROLE NEW	EMMANUELLE\ROUSSOPOULOS\ROUSSOPOULOS\	19	36.2 Mo	4
142	2	PHOTOS POUR DOSSIER CAROLE	EMMANUELLE\ROUSSOPOULOS\ROUSSOPOULOS\	22	506.4 Mo	4
143	2	TEXTE FINAL FILM	EMMANUELLE\ROUSSOPOULOS\ROUSSOPOULOS\	2	117.5 Ko	4

Cet exemple illustre l'existence de « foyers » de redondances ; ces dernières ne sont donc pas réparties de manière totalement aléatoire dans le Fonds Reusser, et cette analyse, facile à modéliser, permet d'attirer l'attention de l'archiviste en charge de l'évaluation sur les dossiers les plus susceptibles de faire l'objet d'une étude détaillée. Cela pourra également être un argument lorsqu'il s'agira de décider du sort final des dossiers redondants : la réunion de ces dossiers au sein d'un même répertoire est-elle le fruit d'une démarche consciente et volontaire de « regroupement » *a posteriori* (justifiant sa conservation) ou ce répertoire a-t-il été éclaté et ses dossiers (re)disséminés ailleurs dans le fonds, en suivant un plan de classification second ? L'étude fine de ce type de « foyers » (par la nomenclature des éléments, les dates de modification ou le voisinage d'autres dossiers non redondants) pourra sans doute livrer de précieux indices.

#### 4.4.3 Comprendre les relations inter-redondances : essais méthodologiques

Les analyses précédentes ont été menées pour pallier les insuffisances des logiciels, qui ne proposent des modalités de traitement qu'au niveau *micro*, par groupes de redondances. Mais ces analyses statistiques ont aussi leur revers : en étudiant les redondances dans leur ensemble, ou en travaillant sur les chemins individuels (comme ci-dessus), elles font trop peu de cas de la manière dont les éléments d'un même groupe sont disséminés au sein des différents répertoires. Il faut tirer parti des avantages de chacun des types d'analyse (les logiciels qui fonctionnent par groupes de redondances et les analyses statistiques qui travaillent sur les chemins et l'ensemble des redondances) et trouver un moyen de conserver la notion de « groupe de redondances » intimement liées par leur contenu tout en mettant en valeur leur distribution dans des répertoires distincts. Pour cela, deux méthodes peuvent être utilisées : une base de données relationnelle et une analyse des embranchements. Les propositions méthodologiques en question sont encore en phase de tests, et nous présentons ci-dessous bien plus une manière possible de travailler que des résultats définitifs à propos des études de cas.

##### 4.4.3.1 Base de données relationnelle entre les répertoires

La base de données relationnelle vise à mettre en avant la manière dont les redondances sont distribuées au sein du répertoire. Le fichier source pour l'établissement de la base possède les colonnes suivantes (voir l'exemple pour les quinze premiers répertoires dans le Tableau 23 : Fichier source pour la base de données relationnelle, dossiers redondants, Fonds Reusser) : à chaque répertoire contenant au moins un dossier redondant est attribué un identifiant unique qui servira à repérer très facilement les répertoires au sein de la base ; le nombre de dossiers redondants par répertoire est indiqué dans une autre colonne ; enfin, les

identifiants du groupe de redondances auquel appartient le dossier redondant que les répertoires hébergent est indiqué dans une colonne, assortie du nom du dossier. Il est alors très facile, par le biais de filtres appliqués à la colonne « ID du groupe de redondances », de savoir quels répertoires entretiennent des liens via le partage de dossiers au contenu commun. À titre d'exemple, le premier répertoire (ID 1, « À CLASSER ») comprend trois dossiers redondants (« LA SUITE DU MONDE », « NOVEMBRE 2012 EXCENEVEVEX » et « paiements à classer ») ; on peut alors filtrer les résultats par les identifiants du groupe de redondances n° 0, 1 et 2, pour isoler les répertoires qui contiennent un ou plusieurs des dossiers en question. Il apparaît alors que le répertoire « À CLASSER » possède deux dossiers communs (ID de groupe 0 et 1) avec le répertoire « À CLASSER\PERSONNEL\ » (ID de répertoire 6) et un élément commun (ID de groupe 2) avec le répertoire « aaCINEATELIER\CINEATELIER 2012\COMPTA 2012\PAIEMENTS EBANKING\ » (ID de répertoire 119).

Tableau 23 : Fichier source pour la base de données relationnelle, dossiers redondants, Fonds Reusser

Répertoire parent		Redondances		
ID de répertoire	Chemin contenant	Nombre de dossiers redondants	ID du groupe de redondances	Nom(s) du dossier redondant
1	A CLASSER\	3	0	LA SUITE DU MONDE
	A CLASSER\		1	NOVEMBRE 2012 EXCENEVEVEX
	A CLASSER\		2	paiements a classer
2	aaCINEATELIER\	1	103	MORAVIA
3	aaREUSSER\	1	103	MORAVIA
4	DIVERS ADRESSES \	3	100	ACTEURS
	DIVERS ADRESSES \		101	CINEMA
	DIVERS ADRESSES \		102	MANU
5	UTILE\	6	115	CODES
	UTILE\		116	CUISINE
	UTILE\		117	HOME
	UTILE\		118	INTERNET
	UTILE\		122	PERSONNEL
6	A CLASSER\PERSONNEL\	2	0	LA SUITE DU MONDE
	A CLASSER\PERSONNEL\		1	NOVEMBRE 2012 EXCENEVEVEX
7	A CLASSER\ROUSSEAU\	2	3	acteurs
	A CLASSER\ROUSSEAU\		4	AFFICHE
8	aaCINEATELIER\CINEATELIER 2007\	1	53	FILMS
9	aaCINEATELIER\CINEATELIER 2008\	2	53	FILMS
10	aaCINEATELIER\CINEATELIER 2009\	2	63	CODES DIVERS
	aaCINEATELIER\CINEATELIER 2009\		71	FILMS
11	aaCINEATELIER\CINEATELIER 2010\		71	FILMS
12	aaCINEATELIER\CINEATELIER 2011\	2	88	DOC_INSTITUTIONS
	aaCINEATELIER\CINEATELIER 2011\		89	FICHIER EMPLOI
13	aaCINEATELIER\CINEATELIER 2012\	3	88	DOC_INSTITUTIONS
	aaCINEATELIER\CINEATELIER 2012\		89	FICHIER EMPLOI
	aaCINEATELIER\CINEATELIER 2012\		98	STATUTS
14	aaCINEATELIER\CINEATELIER 2013\	3	88	DOC_INSTITUTIONS
	aaCINEATELIER\CINEATELIER 2013\		89	FICHIER EMPLOI
	aaCINEATELIER\CINEATELIER 2013\		98	STATUTS
15	aaCINEATELIER\DIVERS ADRESSES \	3	100	ACTEURS
	aaCINEATELIER\DIVERS ADRESSES \		101	CINEMA
	aaCINEATELIER\DIVERS ADRESSES \		102	MANU

On peut alors procéder de la même manière pour les quelque 200 répertoires différents qui contiennent des dossiers redondants. Si ce travail a été réalisé manuellement afin de présenter le concept méthodologique, il va de soi que la base de données pourrait être automatisée moyennant quelques compétences en traitement de données et en programmation informatique. Les résultats sont présentés dans une base de données relationnelle (voir Tableau 25 : « Base de données relationnelle » pour identifier les liens entre répertoires parents, dossiers redondants, Fonds Reusser), avec les identifiants et les noms de répertoires par colonne et par ligne ; le nombre de dossiers aux contenus identiques partagés par deux répertoires est indiqué dans le tableau selon le code couleur ci-dessous :

1	2	3	4	5
---	---	---	---	---

La base de données relationnelle compte en outre les colonnes récapitulatives suivantes :

- **Profondeur** : niveau de profondeur dans l'arborescence du répertoire parent ;
- **Dossiers redondants** : nombre de dossiers non-uniqes contenus dans le répertoire parent, c'est-à-dire le nombre de dossiers qui possèdent au moins une autre occurrence au contenu identique au sein de l'arborescence générale ;
- **Liens avec des répertoires différents** : nombre de relations qu'un répertoire contenant au moins un dossier non-unique entretient avec un répertoire différent contenant un ou plusieurs dossier(s) en commun (un chiffre « 1 » correspond donc à *un lien*, quel que soit le nombre de dossiers que les répertoires partagent) ;
- **Liens totaux** : nombre total de relations qu'un répertoire entretient avec d'autres répertoires pour chacun des dossiers non-uniqes qu'il contient (le chiffre dépend donc du nombre de dossiers non-uniqes et du nombre de liens avec d'autres répertoires) ;

Si les deux premières colonnes (« Profondeur » et « Dossiers redondants ») sont connues et peuvent être remplies sans l'aide de la base de données, les deux colonnes suivantes (« Liens avec des répertoires différents » et « Liens totaux ») sont des champs calculés à partir des valeurs présentes dans la base. Elles sont utiles pour savoir si un répertoire possède des dossiers partagés avec un nombre élevé d'autres répertoires ou si les dossiers redondants qu'il héberge sont communs avec un nombre très restreint d'autres branches de l'arborescence. En d'autres termes, cela indique un « taux de dispersion » plus ou moins élevé des dossiers redondants d'un répertoire : ce dernier entretient-il un lien étroit avec un seul autre répertoire (auquel cas on peut analyser plus finement leurs structures et décider de quelle(s) occurrence(s) on conservera) ou possède-t-il des ramifications multiples (auquel cas, il faudra s'interroger pour savoir s'il s'agit d'un regroupement volontaire d'éléments dispersés, ou s'il s'agit d'un dossier élémentaire qui a fait l'objet d'un « éclatement » second, avec un classement plus fin des éléments dans des répertoires plus détaillés). Le nombre total de liens est une variable qui découle de la première puisqu'il s'agit de la somme totale des valeurs affichées dans le tableau. Cela indique le nombre de liens tissés par chaque répertoire pour tous ses dossiers redondants vers un ou plusieurs autres répertoires. Plus cette valeur est élevée, plus le répertoire contient d'éléments redondants communs avec le reste de l'arborescence.

Pour terminer, on peut croiser les analyses précédentes sur les répertoires parents (voir 4.4.2) et les résultats de la base de données relationnelle pour avoir une vue complète des redondances par dossiers au sein du Fonds Reusser. Nous présentons ci-dessous les résultats pour les répertoires de niveau 2 et 3 contenant des dossiers redondants. À titre d'exemple, le répertoire « À CLASSER » contient en tout 9 dossiers subordonnés, dont 3 sont communs avec d'autres branches de l'arborescence (il y a donc 33 % de dossiers redondants en son sein). Les redondances en question sont distribuées de la manière suivante : 2 dossiers sont communs avec un autre répertoire, et 1 dossier est commun avec un second répertoire (on a donc deux liens avec des répertoires différents, et trois liens en tout). On voit ainsi apparaître de grandes différences dans la manière dont les redondances sont distribuées : le répertoire 5 (« UTILE ») a 6 dossiers en commun avec un seul répertoire (en l'occurrence le répertoire 17, « aaCINEATELIER\UTILE\ » qui a lui-même un dossier redondant avec un autre

répertoire) ; à l'inverse, certains répertoires ne possèdent qu'un seul dossier redondant, mais ce dernier est commun avec de très nombreux autres répertoires.

Tableau 24 : Nombre de dossiers redondants par répertoire de niveau 2 et 3, et mesure de dispersion, Fonds Reusser

Analyse par chemin						Base de données	
ID de répertoire	Répertoire parent	Profondeur	Nbr total de dossiers	Nbr de dossiers redondants	Proportion de dossiers redondants	Liens avec répertoires différents	Liens totaux
1	A CLASSER\	2	9	3	33,33	2	3
2	aaCINEATELIER\	2	44	1	2,27	1	1
3	aaREUSSER\	2	20	1	5,00	1	1
4	DIVERS ADRESSES \	2	5	3	60,00	1	3
5	UTILE\	2	9	6	66,67	1	6
6	A CLASSER\PERSONNEL\	3	2	2	100,00	1	2
7	A CLASSER\ROUSSEAU\	3	6	2	33,33	1	2
8	aaCINEATELIER\CINEATELIER 2007\	3	11	1	9,09	1	1
9	aaCINEATELIER\CINEATELIER 2008\	3	12	2	16,67	2	2
10	aaCINEATELIER\CINEATELIER 2009\	3	14	2	14,29	11	15
11	aaCINEATELIER\CINEATELIER 2010\	3	12	1	8,33	10	14
12	aaCINEATELIER\CINEATELIER 2011\	3	13	2	15,38	2	4
13	aaCINEATELIER\CINEATELIER 2012\	3	15	3	20,00	2	5
14	aaCINEATELIER\CINEATELIER 2013\	3	13	3	23,08	2	5
15	aaCINEATELIER\DIVERS ADRESSES\	3	5	3	60,00	1	3
16	aaCINEATELIER\EN COURS\	3	8	3	37,50	10	12
17	aaCINEATELIER\UTILE\	3	9	7	77,78	2	7
18	UTILE\MAISONS\	3	3	3	100,00	1	3

Cette méthode permet de traiter les redondances de manière hiérarchisée : on débutera le traitement par les répertoires qui contiennent une majorité de dossiers redondants en leur sein (selon le classement de la colonne « Proportion de dossiers redondants ») ; puis, on pourra s'atteler à l'étude des répertoires qui entretiennent des liens avec un nombre restreint de répertoires, les cas étant sans doute plus faciles à trancher ; en dernier lieu, on analysera la composition des répertoires dont le ou les dossiers sont disséminés dans plusieurs branches de l'arborescence. Cette méthode a également l'avantage de faire apparaître quelques « modèles » (*pattern*) dans la manière dont le créateur des documents a utilisé le support de données et dont il a dupliqué les éléments au sein de l'arborescence. À ce stade de l'analyse, et à titre d'exemple, cinq *pattern* ont pu être identifiés (voir 4.4.4 Quels modèles (*pattern*) pour les redondances ?)

Tableau 25 : « Base de données relationnelle » pour identifier les liens entre répertoires parents, dossiers redondants, Fonds Reusser

[illegible]

#### 4.4.3.2 Troncature des chemins

La seconde méthode, qui sera moins développée que la précédente, renverse l'ordre d'analyse : on commence alors par distinguer les groupes de redondances, puis on analyse les répertoires dans lesquels les redondances sont enregistrées (alors que la base de données procédait par répertoires, puis par groupes). Il s'agit alors principalement de mettre ainsi au jour des modèles (*pattern*) dans la manière dont le créateur des documents a utilisé les duplicatas. Pour ce faire, on doit « tronçonner » les chemins complets des répertoires parents, afin de distinguer facilement quels sont les dossiers de niveau 1, de niveau 2, de niveau 3 et ainsi de suite ; l'on peut ainsi, grâce à une mise en forme conditionnelle dans un tableur (Excel en l'occurrence) qui surligne les valeurs non uniques, mettre en avant la structure des répertoires qui contiennent des redondances, et analyser à partir de quel embranchement de l'arborescence les répertoires contenant des redondances divergent. Les exemples ci-dessous proviennent de la recherche de *fichiers* redondants. Nous avons sélectionné quelques cas de figure seulement, les plus évidents sans doute, et une analyse plus poussée mériterait d'être menée, pour vérifier la faisabilité et la généralisation possible de la méthode proposée. Une précision s'impose : les dossiers qui suivent la divergence d'embranchement peuvent certes porter un intitulé identique, mais il ne s'agit pas *stricto sensu* des mêmes éléments (seuls leurs intitulés sont similaires, et une partie seulement de leur contenu – *un fichier au moins* pour que ces dossiers parents apparaissent dans les résultats de recherche de redondances par fichiers, mais non *tous leurs éléments subordonnés*, auquel cas ces dossiers auraient été identifiés comme étant eux-mêmes strictement redondants). Nous pouvons isoler les modèles suivants :

- **l'emboîtement strict**, lorsque tous les intitulés de dossiers sont identiques, mais que seuls les niveaux de profondeur divergent (cela se présente lorsqu'il y a redoublement d'un même nom sur plusieurs niveaux distincts, comme ci-dessous entre les niveaux 4 et 5 avec le dossier « ROUSSOPOULOS ») ;
- **l'enchâssement**, lorsqu'une structure identique pour ce qui est des derniers niveaux est reprise dans un autre répertoire, à un niveau inférieur (décalée de  $n-1$  au minimum) ;
- **les répertoires parallèles**, lorsqu'un embranchement divergent à un niveau intermédiaire de l'arborescence donne ensuite lieu à une structure subordonnée identique (comme c'est le cas ci-dessous pour les dossiers chronologiques, qui reproduisent la même arborescence sous-jacente) ;
- **une arborescence plus détaillée** (de minimum  $n-1$ ), lorsque l'élément redondant se trouve dans la même branche de l'arborescence jusqu'à l'avant-dernier niveau – dans les exemples ci-dessous, cela signifie qu'il y a un élément redondant au niveau  $n$  (dans un dossier final) et un élément redondant identique au niveau  $n+1$  (à la racine d'un dossier non final qui contient au minimum un dossier subordonné dans lequel se trouve la redondance).



Tableau 26 : Analyse des embranchements divergents, fichiers redondants, Fonds Reusser

ID de groupe	Nom	Niveau	Chemin contenant - Niveau 2	Chemin contenant - Niveau 3	Chemin contenant - Niveau 4	Chemin contenant - Niveau 5	Chemin contenant - Niveau 6	Chemin contenant - Niveau 7
<b>EMBOÎTEMENT STRICT</b>								
* La nomenclature des dossiers est identique, mais il y a un décalage dans le niveau de profondeur (un répertoire de même nom est redoublé)								
1698	1 Carole-camera-coBrigitte_new.jpg	5	EMMANUELLE	ROUSSOPOULOS	ROUSSOPOULOS	PHOTOS CAROLE NEW		
1698	1 Carole-camera-coBrigitte_new.jpg	6	EMMANUELLE	ROUSSOPOULOS	ROUSSOPOULOS	ROUSSOPOULOS	PHOTOS CAROLE NEW	
5930	1 Carole-camera-coBrigitte-Pougeoise.	5	EMMANUELLE	ROUSSOPOULOS	ROUSSOPOULOS	PHOTOS POUR DOSSIER CAROLE		
5930	1 Carole-camera-coBrigitte-Pougeoise.	6	EMMANUELLE	ROUSSOPOULOS	ROUSSOPOULOS	ROUSSOPOULOS	PHOTOS POUR DOSSIER CAROLE	
<b>ENCHÂSSEMENT</b>								
* Une structure première et moins profonde est reprise et décalée dans un autre répertoire à un niveau n-1								
5483	adr_alexandra.rtf	3	DIVERS ADRESSES	PERSO				
5483	adr_alexandra.rtf	4	aaCINEATELIER	DIVERS ADRESSES	PERSO			
5080	adresse leo k..eml	3	DIVERS ADRESSES	CINEMA				
5080	adresse leo k..eml	4	aaCINEATELIER	DIVERS ADRESSES	CINEMA			
<b>MULTIPLES RÉPERTOIRES EN PARALLÈLE</b>								
* La distinction se fait à un seul niveau de l'arborescence, avec la reprise des titres de dossiers et de la structure interne.								
* La distinction est typologique, thématique, ou chronologique ou autre								
1447	chatagny.eml	5	aaCINEATELIER	CINEATELIER 2007	FICHER EMPLOI	COMEDIENS		
1447	chatagny.eml	5	aaCINEATELIER	CINEATELIER 2008	FICHER EMPLOI	COMEDIENS		
1447	chatagny.eml	5	aaCINEATELIER	CINEATELIER 2009	FICHER EMPLOI	COMEDIENS		
1447	chatagny.eml	5	aaCINEATELIER	CINEATELIER 2010	FICHER EMPLOI	COMEDIENS		
1447	chatagny.eml	5	aaCINEATELIER	CINEATELIER 2011	FICHER EMPLOI	COMEDIENS		
1447	chatagny.eml	5	aaCINEATELIER	CINEATELIER 2012	FICHER EMPLOI	COMEDIENS		
1447	chatagny.eml	5	aaCINEATELIER	CINEATELIER 2013	FICHER EMPLOI	COMEDIENS		
4163	voyage à chambéry.fdx	6	aaCINEATELIER	aaROUSSEAU	ROUSSEAU_2011	SCENARIO_2011	DIVERS_SCENARIO_FDX Anc	
4163	voyage à chambéry.fdx	6	aaCINEATELIER	aaROUSSEAU	ROUSSEAU_2012	SCENARIO_2011	DIVERS_SCENARIO_FDX Anc	
4024	voyage à Clarens.fdx	6	aaCINEATELIER	aaROUSSEAU	ROUSSEAU_2011	SCENARIO_2011	DIVERS_SCENARIO_FDX Anc	
4024	voyage à Clarens.fdx	6	aaCINEATELIER	aaROUSSEAU	ROUSSEAU_2012	SCENARIO_2011	DIVERS_SCENARIO_FDX Anc	
3538	1.jpg	7	aaCINEATELIER	aaROUSSEAU	ROUSSEAU_2012	AAAIMAGES	ACTEURS	PHOTOS-COMEDIENS
3538	1.jpg	7	aaCINEATELIER	aaROUSSEAU	ROUSSEAU_2012	IMAGES	ACTEURS	PHOTOS-COMEDIENS
<b>ARBORESCENCE PLUS DÉTAILLÉE</b>								
* Si la distinction ne se produit qu'au dernier niveau de l'arborescence, cela signifie qu'il y a un élément à la racine et un élément dans un sous-dossier.								
4783	2 extrait_Dialogue.doc	5	aaCINEATELIER	aaTRINITE_2007_2008	ARCHIVES	TRINITE_2004		
4783	2 extrait_Dialogue.doc	6	aaCINEATELIER	aaTRINITE_2007_2008	ARCHIVES	TRINITE_2004	DOSSIER ANNONCE	
4792	Adresses Larry Page et Sergey Brin.do	6	aaCINEATELIER	ATTALI_1	COEURS	9 LOS ANGELES	8 INTERNET	
4792	Adresses Larry Page et Sergey Brin.do	7	aaCINEATELIER	ATTALI_1	COEURS	9 LOS ANGELES	8 INTERNET	5 GOOGLE

#### 4.4.4 Quels modèles (*pattern*) pour les redondances ?

L'étude des redondances doit, à terme, permettre de mieux comprendre quel a été le fonctionnement du créateur des documents, de manière à pouvoir traiter les dossiers et fichiers en plusieurs exemplaires en toute connaissance de cause et prendre des décisions éclairées quant à leur sort final qui n'engendrent pas de perte d'informations ni du contexte dans lequel les documents ont été créés et sauvegardés. Cette connaissance s'ajoute à celle acquise lors de l'analyse de l'arborescence générale et du plan de classification initial du producteur des archives. Et ce n'est qu'après avoir mené cette enquête, et compris pourquoi et comment le producteur a organisé ses archives et dans quel but des éléments en plusieurs exemplaires ont été générés, que l'on pourra utiliser les fonctionnalités de sélection semi-automatisée proposées par des logiciels comme *AllDup* ou *TreeSize* – en particulier la sélection des éléments redondants par dossiers spécifiques ou par chemins contenant. La reconnaissance de modèles de redondances (*pattern*) propres au producteur de documents *suit* donc l'*identification* des redondances (via les logiciels) et leur *analyse* (via des statistiques générales, une base de données relationnelle ou une étude des embranchements divergents) mais elle *précède* la prise de décision sur le sort qu'on leur réserve et leur traitement effectif (suppression ou conservation).

Dans le Fonds Reusser par exemple, cinq grands modèles (numérotés ci-dessous de 1. à 5.) peuvent être identifiés. Le cas le plus évident concerne les **emboîtements (1.)** de répertoires, à la manière de poupées russes : une arborescence comporte une structure secondaire et subordonnée (quasiment) identique, avec des dossiers aux intitulés et aux contenus similaires. Cela peut concerner une structure complète (qui se répercute de manière identique à un niveau inférieur) ou quelques dossiers seulement.

Figure 45 : Modèles de redondances, emboîtement, Fonds Reusser

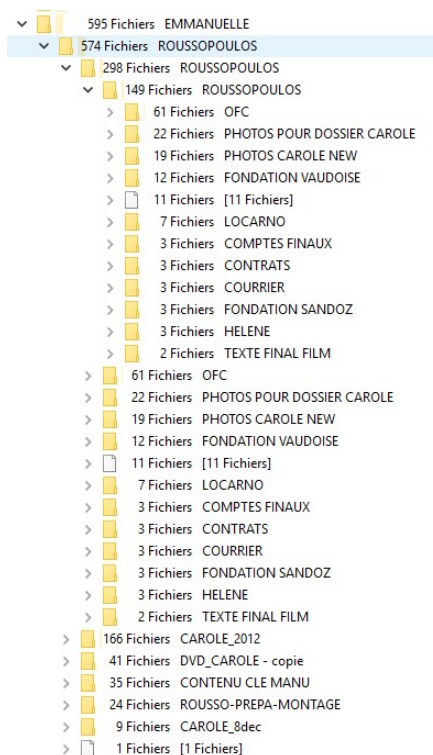
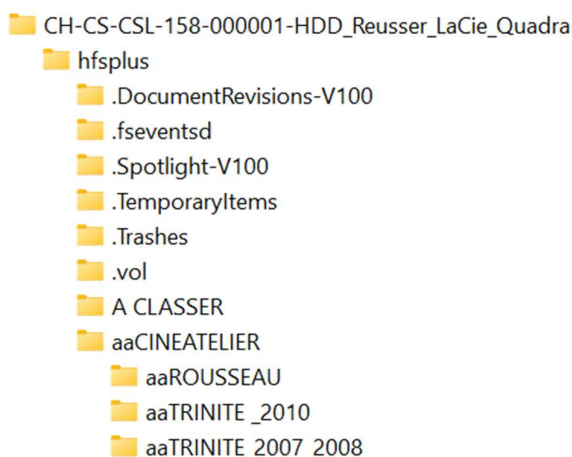


Figure 46 : Modèles de redondances, reports chronologiques, Fonds Reusser

Si cette manière de faire est particulièrement frappante pour les dossiers chronologiques de sa société de production (*Ciné-atelier*), elle est également à l'œuvre dans les répertoires thématiques, liés à un projet qui se prolonge durant plusieurs années. Cela témoigne d'ailleurs d'une certaine pratique du support de données : ce dernier semblait moins être utilisé pour des sauvegardes ponctuelles que pour le travail quotidien. Il s'agissait alors probablement d'avoir « sous la main » tous les dossiers dont on avait (encore) besoin, sans avoir à les chercher dans un répertoire chronologique antérieur ou dans le dossier thématique correspondant (cela explique également la présence des lettres « aa » au début de l'intitulé de certains dossiers – une pratique courante dans le domaine informatique pour voir apparaître au sommet d'une arborescence les dossiers auxquels on a le plus souvent recours).

Figure 47 : Indices de l'utilisation du support : lettres « aa » dans l'intitulé des répertoires, Fonds Reusser



On trouve aussi de très nombreuses redondances dans les dossiers dont l'intitulé témoigne d'un effort de classement, qui a été entrepris, mais dont le processus n'est probablement pas arrivé à son terme – ce que nous pouvons appeler **méta-classement (3.)**. Dès le deuxième niveau de l'arborescence se trouve par exemple un dossier intitulé « À CLASSER », dont on constate via la recherche de redondances que l'immense majorité de ses composants ont en réalité été redistribués dans le fonds, mais que le dossier « original » (si tant est que l'on puisse s'exprimer ainsi) n'a pas été supprimé après l'opération (les dossiers et fichiers ont donc été « copiés » et non « déplacés »). Cela vaut également pour un répertoire intitulé « UTILE » ainsi que pour d'autres sous-dossiers à l'intérieur du Fonds.

On rencontre aussi le cas des **sélections (4.)**, et cela concerne plus souvent les fichiers que les dossiers (les dossiers parents sont ainsi *partiellement* identiques seulement, car un dossier contient une partie, une sélection, de l'ensemble des données contenues dans un répertoire plus généraliste). À titre d'exemple parmi tant d'autres, citons les dossiers de photographies liées au tournage du film de Francis Reusser *Ma Nouvelle Héloïse* (2012) qui contiennent de très nombreux clichés en commun. On trouve ainsi la même photo intitulée « tournage\_reusser02.jpg » dans les dossiers suivants :

- « A CLASSER\ROUSSEAU\IMAGES\Marasco NB\ » ;
- « aaCINEATELIER\aaROUSSEAU\ROUSSEAU\_2012\AAAIMAGES\PHOTOS CHOISIES\MARASCO COMPLET\ » ;
- « aaCINEATELIER\aaROUSSEAU\ROUSSEAU\_2012\IMAGES\PHOTOS\_NH\choix pour francis\ ».

Ou encore, mêlant nom du photographe, date de la prise de vue, et dossier sélectif – avec cette fois, des intitulés différents pour le même cliché, puisque le fichier enregistré dans le dossier sélectif a été renommé :

- « DSC\_2486.JPG » dans le dossier intitulé :  
aaCINEATELIER\aaROUSSEAU\ROUSSEAU\_2012\IMAGES\PHOTOS\_NH\30 MA\DCIM\100D3000\

- « Parapluie.JPG » dans le dossier intitulé :

aaCINEATELIER\aaROUSSEAU\ROUSSEAU\_2012\IMAGES\PHOTOS\_M  
ANU\PHOTOS NH-ROUSSEAU\PHOTOS NOUVELLE HELOÏSE\30  
MA\CHOIX\

Enfin, de manière plus « banale » (mais peu fréquente), on a également affaire à tous les dossiers qui sont en réalité des reliquats d'anciens répertoires, de **sauvegardes antérieures (5.)**, d'arborescence plus anciennes. Ces dossiers sont facilement identifiables car ils contiennent le plus souvent dans leurs titres les mots « copies », « sauvegarde » ou « données anciennes ».

#### 4.4.5 Le travail de sélection des redondances commence

Ces *pattern* mis au jour, charge ensuite à l'archiviste responsable du Fonds Reusser de décider de quelle manière les traiter : conserve-t-on uniquement la première occurrence ou la dernière occurrence d'un élément qui est reporté d'un répertoire chronologique à l'autre ? Conserve-t-on uniquement les dossiers contenant une *sélection* d'éléments (et supprime-t-on, avec ou sans documentation, le dossier qui contient *tous* les fichiers) ? Peut-on éliminer les répertoires de « méta-classement » pour autant que *tous les dossiers* qu'ils contenaient aient bel et bien été reclassés ou est-il « déontologiquement » envisageable, dans le cas contraire, de procéder au (re)classement des éléments non redistribués sans dénaturer le classement initial du producteur ?

Toutes ces questions ont trait non seulement à l'*économie* (ou à l'écosystème) interne du support de données d'un producteur particulier, mais également à la politique de collection générale de l'institution qui conserve ces documents. Quel niveau de traitement désire-t-on appliquer à un fonds d'archives numériques privées : allons-nous établir un plan de classification distinct de celui du producteur ? Est-ce que l'on réorganisera les éléments en fonction d'un plan de classement typologique, chronologique, thématique, mixte ? Quelle profondeur de description choisira-t-on ? Cela dépendra enfin des ressources techniques, humaines et financières à disposition, ou que l'on souhaite investir.

Tableau 27 : Traitement des redondances strictes, tâches

Tâche	Sous-tâches		Méthodes et instruments	Points d'analyse	Identification de modèles (possibles)	Outils
Traitement des redondances strictes	Par Dossiers	Chemin unique		* Checksum	* Nomenclature de l'élément * Date de dernière modification	* Emboîtement (arborescence redoublée) * Reports (chronologiques, thématiques) * Méta-classement * Sauvegarde  * TreeSize Professional (identification)  * Excel (analyse statistique, base de données)  * Tableau (visualisation graphique)
		Chemins multiples	Dossiers avec sous-dossiers	* Checksum * Récolement	* Nomenclature de l'élément * Date de dernière modification	
			Dossiers finaux	* Analyse statistique * Base de données relationnelle * Étude des embranchements	* Niveau de profondeur * Emplacement et voisinage * Liens et organisation intellectuelle	
	Par Fichiers	Chemin unique		* Checksum	* Nomenclature de l'élément * Date de dernière modification * Format de fichier	* Sélection * Compression (formats différents)
		Chemins multiples		* Checksum * Récolement * Analyse statistique * Base de données relationnelle * Étude des embranchements	* Nomenclature de l'élément * Date de dernière modification * Format de fichier * Emplacement et voisinage * Liens et organisation intellectuelle	

## 4.5 Comparaison de données

Dernière étape de la méthodologie que nous proposons pour le tri archivistique des supports de données, la comparaison des fichiers sera menée à des niveaux distincts, qui reprennent la subdivision que nous avons proposée dans la partie consacrée aux outils informatiques (3.3.4), c'est-à-dire la comparaison des répertoires entiers, puis la comparaison des métadonnées des fichiers et, enfin, de manière plus exploratoire à ce stade de l'analyse, la comparaison des images, pour identifier des similarités, via des algorithmes.

### 4.5.1 Comparer des répertoires

La comparaison de répertoires entiers est le prolongement le plus immédiat et le plus évident de la recherche de redondances strictes dont il a été question ci-dessus (la plupart des listes d'outils comme *COPTR* classent d'ailleurs des logiciels comme *Beyond Compare* et *WinMerge* parmi ceux qui permettent de faire de la déduplication). En effet, si nous avons proposé des instruments méthodologiques pour mieux comprendre comment les redondances étaient distribuées dans un fonds d'archives numériques, en analysant tout particulièrement les « chemins parents » ou les répertoires contenant des redondances, afin de pallier l'insuffisance des logiciels qui ne travaillent qu'au niveau des éléments individuels, la comparaison de répertoires ou de dossiers *entiers* constitue la suite logique de notre hypothèse de travail : travailler en premier lieu sur les chemins, afin de toujours conserver à l'esprit l'arborescence du fonds d'archives et la structure interne donnée par le créateur des documents. La comparaison de répertoires et dossiers *entiers*, grâce aux fonctionnalités offertes par *Beyond Compare* permet en effet la visualisation des redondances (éléments strictement identiques) en maintenant la structure de chaque répertoire – tout en faisant aussi apparaître ce qui les distingue ! Quand nous disons que cette comparaison de répertoires est le « prolongement » du traitement des redondances, c'est qu'il est nécessaire de savoir quels répertoires nous désirons analyser (comme on l'a vu pour *Beyond Compare* et *WinMerge* : cette famille de logiciels nécessite que les chemins soient manuellement indiqués aux outils pour qu'ils puissent travailler). Rappelons que le Fonds Reusser compte des milliers de dossiers différents et que les combinaisons entre deux dossiers distincts comparés un à un sont donc extrêmement nombreuses.

La comparaison par répertoires entiers peut s'avérer utile dans plusieurs cas de figure, que ce soit à différents moments de l'acquisition (1), entre différents supports de données provenant de la même acquisition (2), entre des répertoires à la nomenclature proche sur un seul support de stockage (3), ou enfin, si des lots de redondances spécifiques ont été identifiés via des sommes de contrôle à travers plusieurs répertoires d'un même fonds (4) :

1. Tâche utile si l'acquisition de matériel numérique se fait en plusieurs étapes, en plusieurs lots, puisque cela permet de comparer des répertoires différents pour savoir si des fichiers ou des dossiers portant des noms identiques sont présents plusieurs fois, ont déjà été acquis, ou ont été modifiés entre les différents moments de l'acquisition. Cette situation peut se présenter du vivant du donateur (avec des acquisitions successives et des arrivées différées de documents) ou si des documents numériques sont découverts *après* l'acquisition principale.
2. Tâche utile si l'acquisition est composée de plusieurs supports de données qui semblent, via une première évaluation macroscopique, avoir une structure de dossiers identiques (un lot de plusieurs clés USB

par exemple, de plusieurs disques durs externes, ou de plusieurs disquettes).

3. Tâche utile si l'arborescence présente des dossiers aux titres proches, ou aux contenus qui semblent similaires, afin de savoir quel dossier est le plus complet, et comporte le plus grand nombre de fichiers (on s'évite ainsi de devoir comparer les fichiers manuellement pour savoir lesquels sont présents dans un seul dossier ou dans un autre).
4. Tâche utile si l'on reconnaît via un dédoublement des dossiers ou fichiers redondants que plusieurs instances proviennent des mêmes répertoires, ou de dossiers voisins (dossiers qui répondent parfois à des systèmes de nommage différents) : la comparaison des dossiers permet ainsi de voir de manière panoramique quelle est la structure générale du dossier, le nombre de fichiers redondants à l'intérieur des dossiers voisins et, surtout, de comprendre le contexte dans lequel les fichiers redondants ont été créés.

Pour illustrer les bénéfices de l'utilisation d'un logiciel de comparaison de répertoires comme *Beyond Compare*, nous proposons de reprendre quelques-uns des *pattern* précédemment identifiés dans le Fonds Reusser.

Comparer des répertoires dont on a pu identifier via la recherche de redondances strictes qu'il y avait un phénomène d'**emboîtement** est utile pour « contrôler » très rapidement que tous les éléments d'un sous-répertoire qui paraît emboîté dans un registre supérieur sont effectivement identiques. Rappelons que les résultats de recherche des redondances via les logiciels testés fournissent en général le nom du chemin parent directement supérieur, alors que les emboîtements peuvent concerner plusieurs niveaux de répertoires, avec des dossiers qui contiennent des sous-dossiers, comme c'est le cas dans le répertoire :

« EMMANUELLE \ ROUSSOPOULOS \ ROUSSOPOULOS \ ROUSSOPOULOS ».

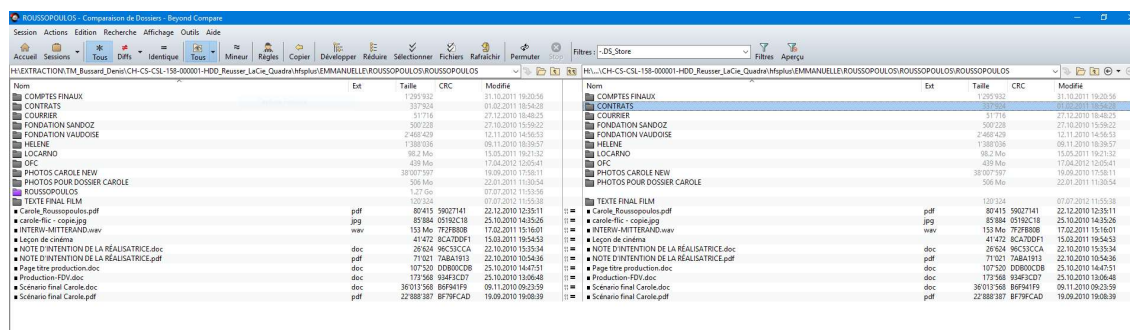
Figure 48 : Structure du répertoire « ROUSSOPOULOS », pour comparaison de dossiers, Fonds Reusser





L'affichage des résultats de recherche de redondances strictes ne permet donc pas de comparer facilement si tous les dossiers et fichiers d'un répertoire *r-1* existent bel et bien au niveau *r* (si l'emboîtement est donc « parfait »). Sans que l'on puisse fournir une explication détaillée, on aperçoit effectivement sur la « Figure 44 : Nombre de dossiers redondants par répertoire parent, Fonds Reusser » et dans le « Tableau 22 : Éléments non-unique au sein d'un répertoire, Fonds Reusser », que le dossier « ROUSSOPOULOS » de niveau inférieur contient en tout 10 dossiers redondants sur 11, et que ces 10 dossiers sont identiques à ceux du répertoire homonyme de niveau supérieur – ce qui signifie que le dossier intitulé « LOCARNO » a bel et bien été reconnu comme identique malgré la présence d'un sous-dossier en son sein (« FICHE-INSCRIPTION ») tandis que le répertoire « OFC » (et ses deux sous-dossiers « OFC2 » et « OFC2 – copie ») n'est pas reconnu comme redondant entre les deux répertoires « ROUSSOPOULOS » emboîtés. Grâce à *Beyond Compare*, on peut donc contrôler que l'emboîtement est parfait, et que l'hypothèse formulée grâce à l'étude des chemins contenant est exacte : les deux répertoires contiennent bien les mêmes dossiers (à l'exception bien sûr du dossier emboîté lui-même « ROUSSOPOULOS », qui apparaît en violet dans le panneau de gauche) :

Figure 49 : Comparaison de répertoires, contrôle d'emboîtement exact, *Beyond Compare*, Fonds Reusser



Le logiciel permet aussi de repérer les différences qui existent entre deux répertoires que l'on a identifiés comme étant proches dans leur contenu. Le cas le plus intéressant concerne les **reports de dossiers dans les répertoires thématiques et chronologiques**. Il s'agit ici de pouvoir analyser plus finement quels sont les dossiers que Francis Reusser a reportés d'un répertoire chronologique à l'autre, c'est-à-dire quel est le degré de similarité entre deux répertoires. Cette information pourra notamment permettre de traiter plus rapidement ces cas si l'on décide de ne garder qu'une seule occurrence de ces répertoires thématico-chronologiques (en supprimant les éléments reportés dans un répertoire plus récent, ou plus significatif, en déplaçant les quelques éléments non reportés si un nouveau plan de classement l'indique, etc.).

L'exemple ci-dessous est une comparaison via *Beyond Compare* des répertoires intitulés « ROUSSEAU\_2011 » et « ROUSSEAU\_2012 », deux dossiers qui documentent la genèse du film *Ma Nouvelle Héloïse* de Francis Reusser, sorti en 2012. Le logiciel ne comparant que les éléments qui possèdent un nom identique, on peut alors « aligner » les éléments qui nous paraissent intéressants pour « forcer » la comparaison. On aperçoit alors non seulement que des dossiers sont reportés entièrement (c'est le cas de « COURRIER » et « COURRIER 2011 », que l'on avait déjà pu identifier via la recherche de redondances strictes), tandis que d'autres ne le sont pas du tout (comme « LES STATIONS DE JÉSUS ») ou alors partiellement (les deux occurrences de « SCÉNARIO\_2011 » par exemple, avec des éléments orphelins de

chaque côté – en violet ; des éléments différents ou plus récents dans le répertoire « ROUSSEAU\_2011 » – en rouge ; et des éléments plus anciens dans le répertoire « ROUSSEAU\_2012 – en gris clair, ce qui peut indiquer que des fichiers ont été modifiés dans le répertoire « ROUSSEAU\_2011 » après le report...).

Figure 50 : Comparaison de répertoires, reports chronologiques, *Beyond Compare*, Fonds Reusser

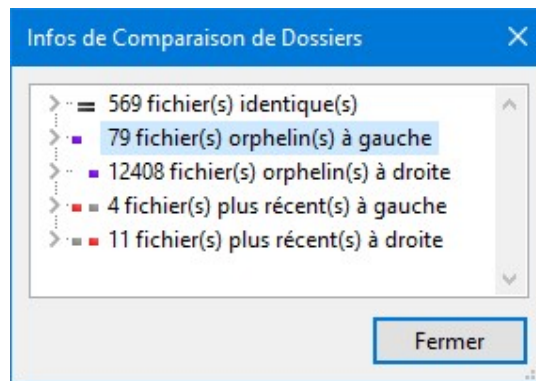
Nom	Taille	CRC	Modifié	Attributs
COMPTA	345 471		28.12.2011 14:05:43	
CONTRATS	5 498 607		16.10.2011 19:25:51	
COURRIER	50 740 948		30.12.2011 19:21:48	
DIVERS_FDX	740 620		28.06.2011 15:34:24	
GALICE	594 363		28.06.2011 15:08:43	
IMAGES	1 104 909 141		28.06.2011 19:32:46	
LES STATIONS DE JESUS	1 539 005		16.01.2011 13:36:15	
NOVEMBRE 2011	17 087 500		15.02.2012 17:38:49	
PRODUCTION ROUSSEAU_2012	1 781 000		15.02.2012 17:55:52	
ROUSSEAU UTILITE	368 731		15.12.2011 11:45:12	
ROUSSEAU EN COURS	1 240 281		24.11.2011 12:54:59	
SCENARIO_2011	4 850 073		05.11.2011 15:14:57	
ACTEURS ROUSSEAU	10 940 243		15.10.2011 12:17:08	
DIVERS SCENARIO	2 519 919		06.10.2011 16:57:12	
DIVERS SCENARIO_FDX_LANC	39 877		27.10.2011 13:41:20	
Fichiers jointe	472 255		26.10.2011 15:15:16	
OCTOBRE	299 384	51D1C3FF	02.11.2011 13:37:39	
MA NOUVELLE HELOISE_éléments pour Aite.doc	27 648	B8673F50	13.09.2011 09:54:47	
MA NOUVELLE HELOISE_éléments pour Aite.pdf	41 218	70F3D54D	13.09.2011 10:25:17	

Les dossiers de **méta-classement** peuvent aussi être traités plus facilement via un logiciel de comparaison de répertoires comme *Beyond Compare*. On peut effectivement utiliser l'affichage des fichiers seulement (en choisissant la fonction « Ignorer la structure de dossier »), sous la forme d'une liste complète des items, quel que soit leur dossier parent, et comparer ainsi *tous les fichiers* d'un répertoire de « méta-classement » avec *tous les éléments* qui figurent dans le répertoire au sein duquel les fichiers ont *probablement* été redistribués. Cela permet de savoir si les fichiers d'un répertoire de méta-classement ont été *totalemt*, *partiellement* ou *pas du tout* reclassés par le créateur des documents. On pourra alors, selon la politique de l'institution, effacer le dossier de méta-classement si l'intégralité de son contenu a été reclassée (ce que l'on aurait également pu voir via la recherche de redondances strictes) ; redistribuer les éléments non reclassés si ces derniers ne sont guère nombreux et qu'un plan de classement archivistique l'impose (en ne travaillant donc que sur les éléments uniques dans le répertoire de méta-classement) ; ou encore conserver les seuls éléments non reclassés si on désire tout à la fois effacer les éléments redondants et conserver la structure et le classement original. Dans l'exemple ci-dessous, on a comparé tous les fichiers du répertoire intitulé « A CLASSER\ROUSSEAU » avec tous les éléments qui figurent dans le dossier « aaCINEATELIER\ROUSSEAU ». Cette vue « à plat », qui ne prend pas en compte la structure de dossiers, permet donc surtout de comparer *tous* les contenus des deux répertoires, et d'identifier les éléments identiques et distincts même si les intitulés de leurs dossiers parents diffèrent.

Figure 51 : Comparaison de répertoires, « méta-classement », éléments identiques (donc reportés), *Beyond Compare*, Fonds Reusser

Figure 52 : Comparaison de répertoires, « méta-classement », éléments orphelins à gauche (non reportés), Fonds Reusser, *Beyond Compare*

Figure 53 : Statistiques de comparaison de dossiers « A CLASSER\ROUSSEAU » et « aaCINEATELIER\ROUSSEAU », *Beyond Compare*, Fonds Reusser



Grâce à la même vue « à plat », *Beyond Compare* permet enfin de comparer les répertoires dont les contenus sont en réalité une **sélection des éléments** présents dans d'autres dossiers (dont seuls quelques items ont donc été dupliqués). On peut ainsi déterminer si les fichiers sélectionnés ont été déplacés (il n'en subsisterait alors qu'une seule instance), ou s'ils ont été copiés (avec donc au minimum deux instances du même document), et surtout avoir une vue panoramique du travail de sélection opéré par le créateur des documents (et non des seuls items dupliqués via une analyse des éléments individuels) : quels sont les éléments copiés ? Combien d'éléments cela représente-t-il par rapport aux éléments totaux ? Y a-t-il une logique à la copie de certains éléments seulement, etc. ? Grâce à l'analyse de l'arborescence, à l'étude de la nomenclature des dossiers et à la recherche de redondances strictes, on est en mesure de déterminer quels sont les dossiers généraux *versus* les dossiers sélectifs que l'on va devoir comparer. Nous reprenons ci-dessous l'exemple des photographies de tournage contenues dans des dossiers qui portent le nom du photographe et / ou la date de la prise de vue :

Figure 54 : Comparaison de répertoires généraux / sélectifs (« MARASCO COMPLET » *versus* « choix pour francis »), *Beyond Compare*, Fonds Reusser

MARASCO COMPLET - choix pour francis - Comparaison de Dossiers - Beyond Compare															
Session Actions Edition Recherche Affichage Outils Aide															
Accueil Sessions Plus récents à gauche, orphelins Identique Agiles Menu Règles Copier Développer Réduire Sélectionner Fichiers Rafraîchir Fenêtre Filtres Agence															
H:\...Mplu\aaCINEATELIER\aaROUSSEAU\ROUSSEAU_2012\AAAMAGES\PHOTOS CHOISIES\MARASCO COMPLET															
				Nom	Chemin	Taille	CRC	Modifié	Attributs						
				Nom	Chemin	Taille	CRC	Modifié	Attributs						
				tournage_reusser108.jpg	...	1978 621	720C726A	12.06.2012 19:56:20	...	tournage_reusser108.jpg	...	1978 621	720C726A	12.06.2012 19:56:20	...
				tournage_reusser107.jpg	...	1544 900	6F1A3294	12.06.2012 19:55:58	...	tournage_reusser107.jpg	...	1544 900	6F1A3294	12.06.2012 19:55:58	...
				tournage_reusser106.jpg	...	1620 088	1C210457	12.06.2012 19:55:38	...	tournage_reusser106.jpg	...	1620 088	1C210457	12.06.2012 19:55:38	...
				tournage_reusser105.jpg	...	1664 037	630B7C73	12.06.2012 19:54:54	...	tournage_reusser105.jpg	...	1664 037	630B7C73	12.06.2012 19:54:54	...
				tournage_reusser104.jpg	...	1741 811	26A47BFB	12.06.2012 19:54:26	...	tournage_reusser104.jpg	...	1741 811	26A47BFB	12.06.2012 19:54:26	...
				tournage_reusser103.jpg	...	1467 330	D1059F03	12.06.2012 19:54:12	...	tournage_reusser103.jpg	...	1467 330	D1059F03	12.06.2012 19:54:12	...
				tournage_reusser102.jpg	...	9	00000000	04.06.2012 23:01:12	...	tournage_reusser102.jpg	...	9	00000000	04.06.2012 23:01:12	...
				tournage_reusser101.jpg	...	1152 394	D1537698	04.06.2012 23:00:50	...	tournage_reusser101.jpg	...	1152 394	D1537698	04.06.2012 23:00:50	...
				tournage_reusser79.jpg	...	1393 406	97E3A47	04.06.2012 23:00:32	...	tournage_reusser79.jpg	...	1393 406	97E3A47	04.06.2012 23:00:32	...
				tournage_reusser78.jpg	...	1264 319	F1BEFC09	04.06.2012 22:59:56	...	tournage_reusser78.jpg	...	1264 319	F1BEFC09	04.06.2012 22:59:56	...
				tournage_reusser77.jpg	...	1668 480	45191919	02.06.2012 20:13:12	...	tournage_reusser77.jpg	...	1668 480	45191919	02.06.2012 20:13:12	...
				tournage_reusser76.jpg	...	1837 729	0A6B467	02.06.2012 21:14:54	...	tournage_reusser76.jpg	...	1837 729	0A6B467	02.06.2012 21:14:54	...
				tournage_reusser75.jpg	...	1660 320	0A0D792	02.06.2012 20:11:54	...	tournage_reusser75.jpg	...	1660 320	0A0D792	02.06.2012 20:11:54	...
				tournage_reusser74.jpg	...	2153 334	57A6444	02.06.2012 20:10:44	...	tournage_reusser74.jpg	...	2153 334	57A6444	02.06.2012 20:10:44	...
				tournage_reusser73.jpg	...	955 066	4C14BA47	01.06.2012 00:48:44	...	tournage_reusser73.jpg	...	955 066	4C14BA47	01.06.2012 00:48:44	...
				tournage_reusser72.jpg	...	1902 629	CC25744	02.06.2012 20:08:36	...	tournage_reusser72.jpg	...	1902 629	CC25744	02.06.2012 20:08:36	...
				tournage_reusser71.jpg	...	1918 186	1F91A4E	02.06.2012 20:08:16	...	tournage_reusser71.jpg	...	1918 186	1F91A4E	02.06.2012 20:08:16	...
				tournage_reusser70.jpg	...	2018 807	3E3768AA	02.06.2012 20:07:56	...	tournage_reusser70.jpg	...	2018 807	3E3768AA	02.06.2012 20:07:56	...
				tournage_reusser69.jpg	...	1429 702	ACF0F888	02.06.2012 20:07:30	...	tournage_reusser69.jpg	...	1429 702	ACF0F888	02.06.2012 20:07:30	...
				tournage_reusser68.jpg	...	1228 934	589AA7B	02.06.2012 20:07:02	...	tournage_reusser68.jpg	...	1228 934	589AA7B	02.06.2012 20:07:02	...
				tournage_reusser67.jpg	...	2014 865	8B16999	02.06.2012 20:06:46	...	tournage_reusser67.jpg	...	2014 865	8B16999	02.06.2012 20:06:46	...
				tournage_reusser66.jpg	...	2318 244	99AB3C70	02.06.2012 20:06:32	...	tournage_reusser66.jpg	...	2318 244	99AB3C70	02.06.2012 20:06:32	...
				tournage_reusser65.jpg	...	2238 195	8A99ACE	02.06.2012 20:06:08	...	tournage_reusser65.jpg	...	2238 195	8A99ACE	02.06.2012 20:06:08	...
				tournage_reusser64.jpg	...	2301 770	71A3EAD	02.06.2012 20:05:52	...	tournage_reusser64.jpg	...	2301 770	71A3EAD	02.06.2012 20:05:52	...
				tournage_reusser63.jpg	...	2031 563	85A4A3E	02.06.2012 20:04:40	...	tournage_reusser63.jpg	...	2031 563	85A4A3E	02.06.2012 20:04:40	...
				tournage_reusser62.jpg	...	2050 989	A6B1293	02.06.2012 20:04:22	...	tournage_reusser62.jpg	...	2050 989	A6B1293	02.06.2012 20:04:22	...
				tournage_reusser61.jpg	...	1706 307	A1E1F38B	02.06.2012 20:03:24	...	tournage_reusser61.jpg	...	1706 307	A1E1F38B	02.06.2012 20:03:24	...
				tournage_reusser60.jpg	...	833 139	DE0DA0C9	02.06.2012 20:01:56	...	tournage_reusser60.jpg	...	833 139	DE0DA0C9	02.06.2012 20:01:56	...
				tournage_reusser59.jpg	...	1711 888	04F820D7	02.06.2012 20:01:34	...	tournage_reusser59.jpg	...	1711 888	04F820D7	02.06.2012 20:01:34	...
				tournage_reusser58.jpg	...	1444 707	DDF7B44	02.06.2012 20:01:14	...	tournage_reusser58.jpg	...	1444 707	DDF7B44	02.06.2012 20:01:14	...
				tournage_reusser57.jpg	...	1950 369	019DCA4	02.06.2012 20:00:40	...	tournage_reusser57.jpg	...	1950 369	019DCA4	02.06.2012 20:00:40	...
				tournage_reusser56.jpg	...	1549 286	CACBA7C8	02.06.2012 19:59:44	...	tournage_reusser56.jpg	...	1549 286	CACBA7C8	02.06.2012 19:59:44	...
				tournage_reusser55.jpg	...	1537 689	498B3A03	02.06.2012 19:59:24	...	tournage_reusser55.jpg	...	1537 689	498B3A03	02.06.2012 19:59:24	...
				tournage_reusser54.jpg	...	1431 724	FC11822	02.06.2012 19:59:04	...	tournage_reusser54.jpg	...	1431 724	FC11822	02.06.2012 19:59:04	...
				tournage_reusser53.jpg	...	1734 873	B8E7C32	02.06.2012 19:58:42	...	tournage_reusser53.jpg	...	1734 873	B8E7C32	02.06.2012 19:58:42	...
				tournage_reusser52.jpg	...	1432 815	295A6C05	02.06.2012 19:57:50	...	tournage_reusser52.jpg	...	1432 815	295A6C05	02.06.2012 19:57:50	...
				tournage_reusser51.jpg	...	1666 701	19A14440	02.06.2012 19:54:30	...	tournage_reusser51.jpg	...	1666 701	19A14440	02.06.2012 19:54:30	...
				tournage_reusser50.jpg	...	1879 710	9A36A80	02.06.2012 19:54:14	...	tournage_reusser50.jpg	...	1879 710	9A36A80	02.06.2012 19:54:14	...
				tournage_reusser49.jpg	...	1460 998	154B1E18	02.06.2012 19:53:20	...	tournage_reusser49.jpg	...	1460 998	154B1E18	02.06.2012 19:53:20	...
				tournage_reusser48.jpg	...	1286 238	93517C93	02.06.2012 19:53:06	...	tournage_reusser48.jpg	...	1286 238	93517C93	02.06.2012 19:53:06	...
				tournage_reusser47.jpg	...	993 144	4F6B467E	02.06.2012 19:51:32	...	tournage_reusser47.jpg	...	993 144	4F6B467E	02.06.2012 19:51:32	...
				tournage_reusser46.jpg	...	317 456	F1ED0E2	03.06.2012 13:57:57	...	tournage_reusser46.jpg	...	317 456	F1ED0E2	03.06.2012 13:57:57	...
				tournage_reusser45.jpg	...	1879 596	SCA14B4A	02.06.2012 19:51:31	...	tournage_reusser45.jpg	...	1879 596	SCA14B4A	02.06.2012 19:51:31	...
				tournage_reusser44.jpg	...	456 396	D1C5324	02.06.2012 13:57:57	...	tournage_reusser44.jpg	...	456 396	D1C5324	02.06.2012 13:57:57	...
				tournage_reusser43.jpg	...	474 607	122AB4B8	02.06.2012 13:57:57	...	tournage_reusser43.jpg	...	474 607	122AB4B8	02.06.2012 13:57:57	...
				tournage_reusser42.jpg	...	555 560	7A46492	02.06.2012 13:57:57	...	tournage_reusser42.jpg	...	555 560	7A46492	02.06.2012 13:57:57	...
				tournage_reusser41.jpg	...	960 203	B0B17D81	02.06.2012 19:50:24	...	tournage_reusser41.jpg	...	960 203	B0B17D81	02.06.2012 19:50:24	...
				tournage_reusser40.jpg	...	475 312	4A43C424	02.06.2012 13:57:57	...	tournage_reusser40.jpg	...	475 312	4A43C424	02.06.2012 13:57:57	...
				tournage_reusser39.jpg	...	269 045	86F0E086	02.06.2012 13:57:57	...	tournage_reusser39.jpg	...	269 045	86F0E086	02.06.2012 13:57:57	...
				tournage_reusser38.jpg	...	1962 395	CDE00774	02.06.2012 19:48:02	...	tournage_reusser38.jpg	...	1962 395	CDE00774	02.06.2012 19:48:02	...
				tournage_reusser37.jpg	...	282 965	82518819	02.06.2012 13:57:57	...	tournage_reusser37.jpg	...	282 965	82518819	02.06.2012 13:57:57	...
				tournage_reusser36.jpg	...	884 078	E22D4415	02.06.2012 19:47:42	...	tournage_reusser36.jpg	...	884 078	E22D4415	02.06.2012 19:47:42	...
				tournage_reusser35.jpg	...	278 296	97C26A48	02.06.2012 13:57:57	...	tournage_reusser35.jpg	...	278 296	97C26A48	02.06.2012 13:57:57	...
				tournage_reusser34.jpg	...	481 081	E451D508	02.06.2012 13:57:57	...	tournage_reusser34.jpg	...	481 081	E451D508	02.06.2012 13:57:57	...
				tournage_reusser33.jpg	...	1412 932	4A8F39C3	02.06.2012 19:47:06	...	tournage_reusser33.jpg	...	1412 932	4A8F39C3	02.06.2012 19:47:06	...



## 4.5.2 Comparer les métadonnées des fichiers

La recherche de redondances strictes, au niveau des fichiers, fournissait notamment les résultats suivants pour le Fonds Reusser : 6'050 lots de fichiers redondants, dont seuls 30 (!) groupes contenaient des fichiers multiples avec des extensions différentes. Si l'on analyse les données dans le détail, on aperçoit cependant que la majorité des différences d'extension proviennent soit de l'absence d'extension pour un fichier redondant au moins (comme pour le groupe portant l'identifiant 4100 ci-dessous), soit en raison d'extensions qui n'expriment pas un format de fichier particulier : sachant que le logiciel considère comme une extension tous les caractères alphanumériques qui suivent le dernier point du titre, on voit apparaître dans la colonne « Extension » des informations relatives au fichier ou à son statut – « .tif NB copie » par exemple dans le groupe numéroté 3833, ou « .3 » et « .4 » pour la numérotation d'un scénario (groupe n° 3869).

Tableau 28 : Exemples de fichiers redondants de différentes extensions, Fonds Reusser

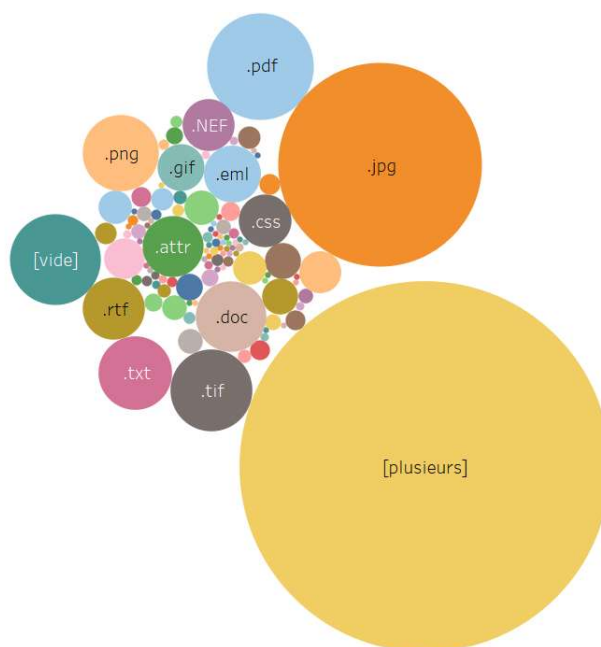
ID de groupe	Nom	Chemin contenant	Extension	Fichiers	Taille	Dernière modification	Type	Niveau de répertoire
	CF Ramuz.tif NB	[plusieurs]	[plusieurs]	2	212.1 Ko	[plusieurs]	Fichier	[plusieurs]
3833	CF Ramuz.tif NB	hfspplusaaCINEATELIER\COUV_DVD\DVD derborene\	.tif NB	1	106.1 Ko	08.09.2004	Fichier	4
3833	CF Ramuz.tif NB copie	hfspplusaaCINEATELIER\RAMUZ\IMAGES RAMUZ DIVERS\IMAGES FILMS\	.tif NB copie	1	106.1 Ko	08.09.2004	Fichier	5
	ramuz RAPT - fille.tif	[plusieurs]	[plusieurs]	4	453.9 Ko	[plusieurs]	[plusieurs]	[plusieurs]
3816	ramuz RAPT - fille	hfspplusaaCINEATELIER\RAMUZ\CORINNA BILLE RAPT\IMAGES RAPT\		1	113.5 Ko	02.10.2004	Fichier	5
3816	ramuz RAPT - fille.tif	hfspplusaaCINEATELIER\RAMUZ\CORINNA BILLE RAPT\IMAGES RAPT\	.tif	1	113.5 Ko	02.10.2004	Fichier TIFF	5
3816	ramuz RAPT - fille.tif	hfspplusaaCINEATELIER\RAMUZ\IMAGES RAMUZ DIVERS\IMAGES FILMS\	.tif	1	113.5 Ko	02.10.2004	Fichier TIFF	5
3816	ramuz RAPT - fille	hfspplusaaCINEATELIER\RAMUZ\IMAGES RAMUZ DIVERS\IMAGES FILMS\IMAGES RAPT\		1	113.5 Ko	02.10.2004	Fichier	6
	Le Cheval Frontière	hfspplusaaCINEATELIER\TV 2001\JURA\COURRIER\MAILS JURA\	[plusieurs]	2	12.3 Ko	[plusieurs]	[plusieurs]	6
5125	Cheval sandoz 31.07.01	hfspplusaaCINEATELIER\TV 2001\JURA\COURRIER\MAILS JURA\	.01	1	6.1 Ko	06.08.2001	Fichier 01	6
5125	Le Cheval Frontière	hfspplusaaCINEATELIER\TV 2001\JURA\COURRIER\MAILS JURA\		1	6.1 Ko	04.08.2001	Fichier	6
	Notre vie	[plusieurs]	[plusieurs]	2	125.3 Ko	28.01.2006	[plusieurs]	[plusieurs]
4100	camera_lac.jpg	hfspplusaaCINEATELIER\SITE_CA\	.jpg	1	62.7 Ko	28.01.2006	Fichier JPG	3
4100	Notre vie	hfspplusaaCINEATELIER\SITE_CA\SITE_ELEMENT\IMAGES SITE CA\		1	62.7 Ko	28.01.2006	Fichier	5
	solgatsquerre.tif	[plusieurs]	[plusieurs]	3	33.4 Mo	[plusieurs]	[plusieurs]	[plusieurs]
306	solgatsquerre.tif	hfspplusaaCINEATELIER\COUV_DVD\DVD Guerre\	.tif	1	11.1 Mo	08.09.2004	Fichier TIF	4
306	solgats querre.tif copie	hfspplusaaCINEATELIER\RAMUZ\IMAGES RAMUZ DIVERS\	.tif copie	1	11.1 Mo	08.09.2004	Fichier	4
306	solgatsquerre.tif	hfspplusaaCINEATELIER\RAMUZ\IMAGES RAMUZ DIVERS\IMAGES FILMS\	.tif	1	11.1 Mo	08.09.2004	Fichier TIF	5
	Scénario 8.4	hfspplusaaREUSSER\ANCIEN 87-96\DONNEES ANCIENNES\CINEMA 96\ENFIN LIBRE\	[plusieurs]	2	192.0 Ko	17.12.1995	[plusieurs]	6
3869	Scénario 8.3	hfspplusaaREUSSER\ANCIEN 87-96\DONNEES ANCIENNES\CINEMA 96\ENFIN LIBRE\	.3	1	96.0 Ko	17.12.1995	Fichier 3	6
3869	Scénario 8.4	hfspplusaaREUSSER\ANCIEN 87-96\DONNEES ANCIENNES\CINEMA 96\ENFIN LIBRE\	.4	1	96.0 Ko	17.12.1995	Fichier 4	6

Un parcours sommaire de l'arborescence avait pourtant laissé entrevoir une pratique fréquente de la part de Francis Reusser : la multiplication de fichiers aux intitulés identiques encodés dans des formats différents, que ce soit dans des formats JPEG (« Joint Photographic Experts Group », un format d'enregistrement et de compression d'images fixes) et NEF (« Nikon Digital SLR Camera Raw Image File », un format propre à Nikon qui enregistre les images de manière non compressée ou avec une compression sans perte) ou dans des formats DOC (« Microsoft Word Document ») et PDF (« Acrobat PDF – Portable Document Format »). L'encodage de ces contenus ainsi que le taux de compression (avec des poids de fichiers extrêmement différents) rendent la recherche de redondances via des sommes de contrôle caduque, alors même que les intitulés de fichiers sont identiques (et que leur contenu semble visuellement et intellectuellement similaire selon les premiers pointages manuels qui peuvent être effectués à travers le fonds).

Il peut donc s'avérer intéressant de mener une recherche de fichiers possédant des métadonnées communes. Cette recherche mérite d'être entreprise seulement *après* l'identification des redondances strictes (via les sommes de contrôle) car certains fichiers aux caractéristiques communes ont peut-être déjà été traités via les valeurs de hachage (ce qui réduit donc le nombre de résultats à analyser dans le cadre de la seconde recherche) et parce que l'analyse des items qui partagent des métadonnées identiques est bien plus détaillée et chronophage.

*TreeSize* et *WinCatalog* proposent notamment l'identification de fichiers aux métadonnées identiques via leur onglet de recherche consacré aux « doublons ». Sachant que nous désirons élucider le cas des fichiers aux intitulés identiques mais aux formats d'encodage distincts, nous privilégierons *TreeSize* qui offre la possibilité de lancer une recherche sur les noms de fichiers, mais surtout, sur les noms de fichiers sans extension<sup>19</sup>. De cette manière, un fichier « .doc » et un fichier « .pdf » qui partagent le même titre apparaîtront dans un même onglet. Les cas de figure sont d'ailleurs extrêmement fréquents dans le Fonds Reusser (rappelons une fois encore que la recherche a été menée dans le fonds entier, expurgé ni de ses fichiers système ni de ses redondances strictes, qui sont donc incluses dans les chiffres suivants). Via la recherche « Nom sans extension » (le logiciel compare alors les caractères qui précèdent le dernier « . » qui figure dans le titre), *TreeSize* a identifié 28'893 objets (représentant 178.5 Go), répartis dans 7'247 groupes, qui partagent des titres identiques, sans prendre en compte leur extension. Si on s'intéresse maintenant aux fichiers de même nom mais qui possèdent des extensions différentes (objet de notre présente étude), on obtient un nombre très conséquent de groupes avec 3'813 lots (soit plus de la moitié, avec 52 %). En d'autres termes, il y a 3'813 intitulés identiques pour des fichiers (deux au minimum) qui sont encodés dans plusieurs formats différents (représentés ci-dessous par la pastille jaune, « [plusieurs] »).

Figure 55 : Proportion des intitulés identiques de fichiers, par extension, Fonds Reusser



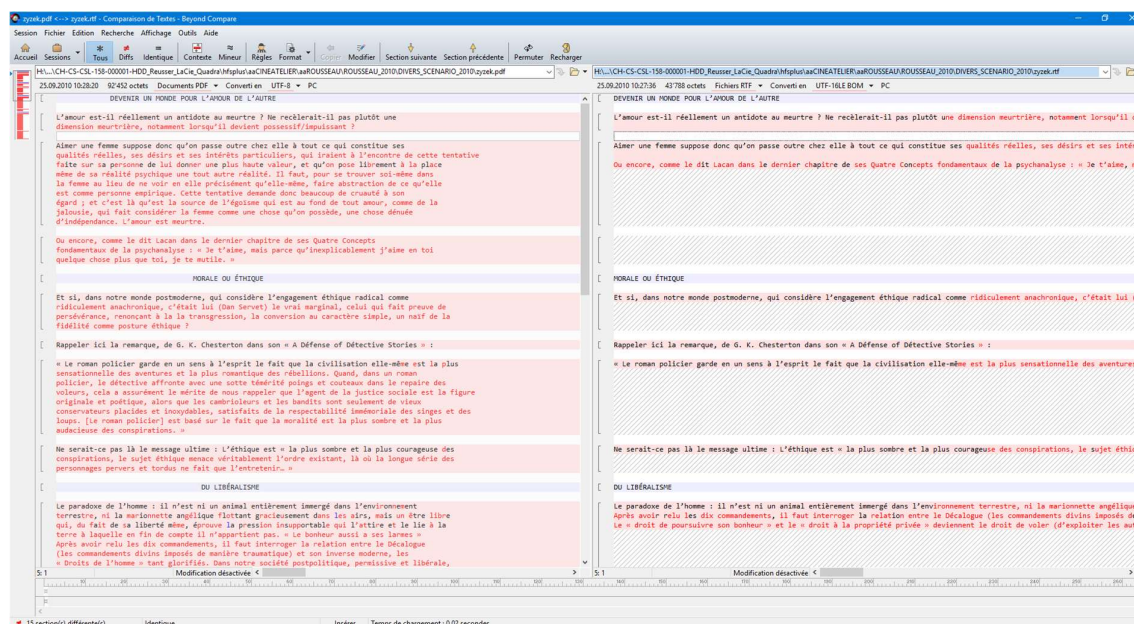
Le traitement des résultats de la comparaison de métadonnées dépassant de loin le cadre de cette étude, nous proposons ci-dessous un exemple tiré de la coexistence, dans le Fonds Reusser, de multiples groupes de fichiers aux titres identiques mais encodés en format « .doc » et « .pdf ». Si l'intuition initiale lors du parcours sommaire du Fonds (qui consiste à

<sup>19</sup> Le logiciel présente cette option de la manière suivante (il met d'ailleurs en avant la possibilité de trouver des images qui possèdent des taux de compression différents) : « Select this option to detect files with equal names, without regarding the file extension. This can be interesting in case you are searching for duplicated backup files or e.g. row-data and compact image or video files ("MyPhoto.bmp" - "MyPhoto.png") » (Jam Software (Joachim Marder) 2022, p. 114)



l'archiviste de savoir avec exactitude si les contenus sont intellectuellement et strictement les mêmes. *Beyond Compare* offre ainsi la possibilité de comparer deux fichiers textuels, même si l'un d'eux est encodé en format « .pdf » (ce que ne permet pas *WinMerge*). Le résultat s'affiche alors sous la forme de deux panneaux, dans lesquels figurent les fichiers comparés, avec la mise en valeur de leurs différences et ressemblances selon un code couleur propre à l'outil. On constate alors, lorsqu'on indique au logiciel de ne pas prendre en compte les différences mineures (des modifications d'espaces blancs par exemple), que les deux fichiers contiennent le même texte, d'un point de vue intellectuel.

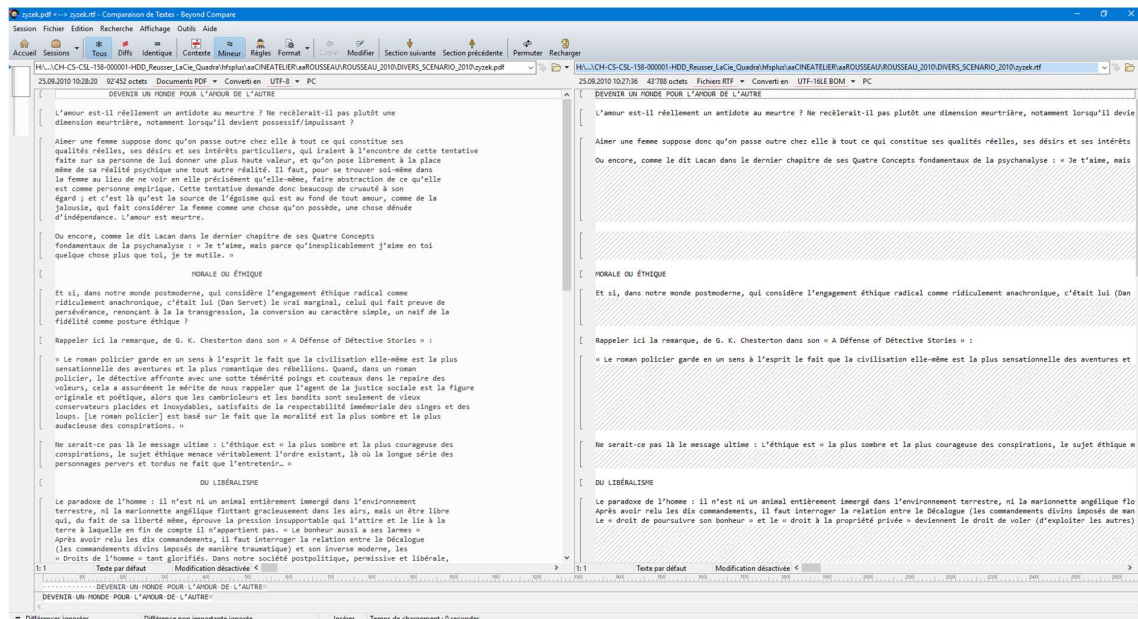
Figure 57 : Comparaison de fichiers texte, avec toutes les différences, *Beyond Compare*, Fonds Reusser



expérimentaux permettant de faire réapparaître à l'écran les états successifs d'une rédaction : ils donnent à voir la concaténation temporelle des métamorphoses du texte, séquence par séquence, chaque état variant ayant été indexé et sauvegardé » (Biasi 2013, p. 43).



Figure 58 : Comparaison de fichiers texte, différences mineurs ignorées, *Beyond Compare*, Fonds Reusser



Une telle comparaison reste évidemment un cas à part, dans le cadre d'une évaluation *macro* adoptant les principes *MPLP (More Product, Less Process)* (Greene, Meissner 2005) : il n'est ni souhaitable, ni envisageable de comparer les fichiers texte deux à deux, comme le propose *Beyond Compare* (ou trois à trois via *WinMerge*) pour un fonds contenant plus de 38'000 fichiers. En revanche, une telle analyse, via des logiciels faciles d'utilisation comme ceux présentés dans cette étude, permet de repérer des pratiques courantes (des *pattern*) dans le fonctionnement interne d'un fonds d'archives numériques privées. On peut effectuer ce type de tests sur un échantillon seulement des fichiers portant le même titre, et voir alors si cela traduit un fonctionnement récurrent dont on pourrait tirer des enseignements pour le tri de données massives.

#### 4.5.3 Comparaison d'images via des algorithmes

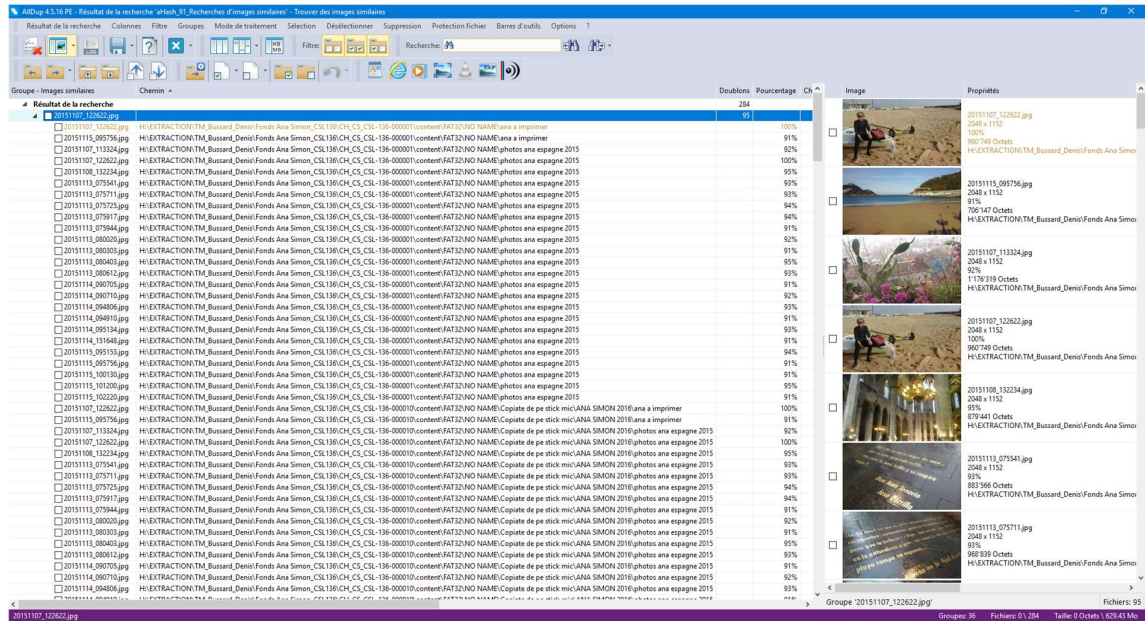
Une étude de cas de grande ampleur relative aux comparaisons d'images similaires n'est malheureusement pas possible dans le cadre de cette recherche. Il importe toutefois de souligner que cette étape intervient à la fin de la méthodologie de tri archivistique au niveau *macroscopique*. À l'instar de ce que nous venons d'évoquer concernant la comparaison fine de deux fichiers texte suite à l'identification de métadonnées communes, la comparaison d'images similaires est à la frontière entre plusieurs champs de l'évaluation : si jusque-là toutes les étapes portaient quasi exclusivement sur les métadonnées des fichiers ou les sommes de contrôle, la comparaison des blocs et lignes de fichiers texte et la confrontation de deux images similaires touchent à la difficile question de l'évaluation des *contenus*. Car si des logiciels permettent effectivement de faciliter l'identification des redondances et des similarités, dans les deux cas (textes légèrement différents ou images similaires mais non identiques), il s'agit par la suite, une fois que les résultats de l'analyse s'affichent à l'écran, de comparer visuellement et intellectuellement les contenus pour décider du fichier à conserver en priorité. En outre, les problématiques abordées lors du traitement des redondances strictes demeurent : quel fichier doit-on privilégier en fonction du répertoire dans lequel il se trouve, si les dossiers parents ne sont pas les mêmes. Avec la comparaison de fichiers au contenu

(« seulement ») similaire, les questions ne cessent de se multiplier : quel fichier garde-t-on si les contenus sont similaires mais non identiques ? Et quelle occurrence conserve-t-on en fonction de son contexte d'enregistrement ?

À ces questions générales s'ajoutent des difficultés d'ordre technique : les deux logiciels testés (*AllDup* et *AntiDupl*) proposent des algorithmes différents, et il est possible, pour chacun des algorithmes de comparaison, de choisir la valeur seuil à partir de laquelle deux images sont reconnues par le logiciel comme étant similaires (et méritant ainsi d'être affichées dans les résultats de recherche). Sachant que *AntiDupl* propose deux algorithmes et qu'*AllDup* en propose quatre (*Average Hash*, *BlockHash*, *Difference Hash* et *Perceptual Hash*) et que le seuil de tolérance s'exprime en pourcentage, cela donne une idée du nombre de combinaisons possibles... – ce qui décourage toutes velléités d'établir des statistiques sur les niveaux de performance des logiciels et sur le nombre d' « images similaires » que contient le Fonds Reusser. Il existe cependant des articles et des blogs qui ont procédé à des comparaisons, en travaillant sur des échantillons de photographies pour calculer le nombre de faux positifs – citons à titre d'exemple le comparatif réalisé par *content-blockchain.org* qui recommande l'utilisation du *Perceptual Hash* (Content Blockchain 2019).

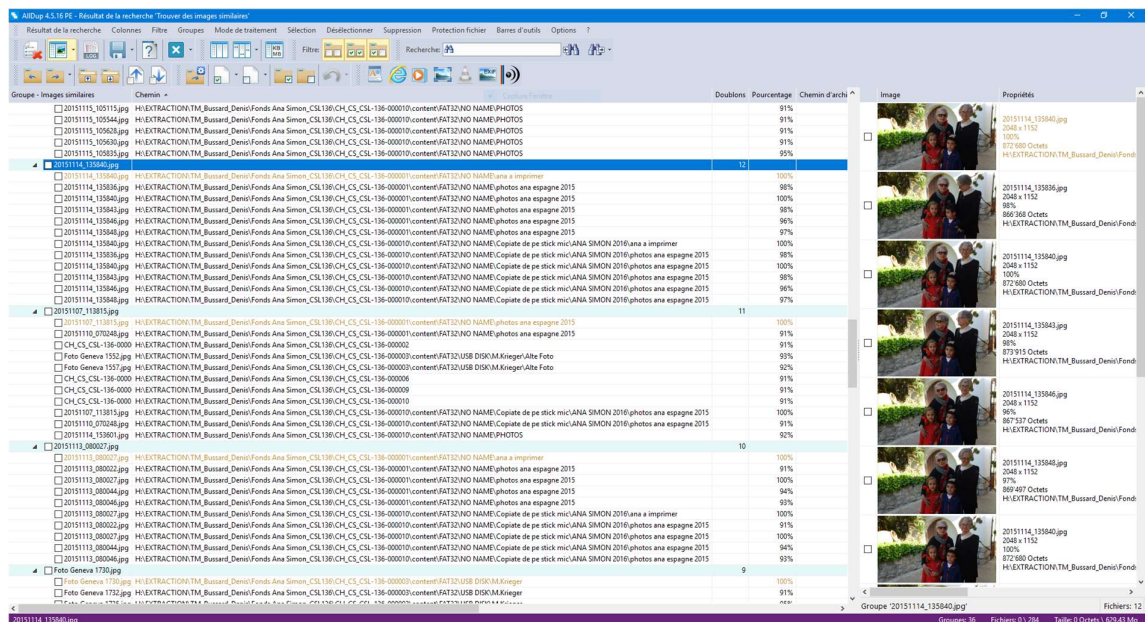
Pour illustrer le fonctionnement du logiciel *AllDup*, et la difficulté de tirer des conclusions générales pour l'ensemble des fonds d'archives privées, nous pourrions donner les deux exemples ci-dessous. La recherche d'images similaires (via l'algorithme de comparaison *Average Hash* et un pourcentage de concordance de 91 %, soit les valeurs indiquées par défaut par le logiciel) dans le Fonds Ana Simon (qui contient en tout quelque 1'200 fichiers dont plus de 300 images) ayant donné lieu à la création de 937 sommes de contrôle et à plus de 47'000 comparaisons *inter-fichiers* en moins de 4 minutes a livré les résultats suivants : 284 « doublons » répartis dans 36 « groupes ». Parmi les « groupes », l'un d'eux contient 95 fichiers et les résultats sont pour le moins troublants comme on peut le voir sur la figure ci-dessous (le fichier dont le nom est en couleur constitue le fichier de « référence » pour la comparaison selon les maigres explications données par le logiciel : « The calculated percentage matching between two pictures will be shown at the column *Match*. The percentage always refers to the reference picture of a group which will be shown in a different text color » (MTSD 2022)).

Figure 59 : Comparaison d'images similaires, exemple d'échec, *AllDup*, Fonds Simon



A *contrario*, le groupe de « doublons » ci-dessous présente des résultats bien plus satisfaisants : on y reconnaît facilement les photographies identiques (les « vraies » redondances, c'est-à-dire les fichiers dont il existe deux instances dans des répertoires différents), mais également des clichés similaires puisque seule l'orientation du visage de Ana Simon diffère d'une image à l'autre. On a bien affaire à une série de photographies, dont il n'est peut-être pas nécessaire de conserver toutes les manifestations (mais la question demeure : comment trancher ? Selon quel(s) critère(s) et avec quelle documentation ?).

Figure 60 : Comparaison d'images similaires, exemple de réussite, *AllDup*, Fonds Simon



Avant de procéder au traitement des fichiers similaires, il faudra donc faire quelques pointages dans le fonds d'archives pour voir si la photographie en rafales, en série, est une pratique courante, ou si un tri avait déjà été opéré ; si une telle « déduplication » se justifie, alors il convient de faire des essais avec les logiciels cités, les algorithmes proposés et les différents seuils de tolérance ; enfin, en fonction des résultats, on pourra s'interroger sur la suite à donner au traitement : quelle est la pratique du créateur des documents (est-ce habituel, ponctuel) ? Les photographies similaires témoignent-elles d'une démarche consciente et volontaire (Francis Reusser ayant suivi une formation de photographe, on se montrera par exemple plus scrupuleux et précautionneux dans le traitement de photographies que l'on pourrait un peu rapidement considérer comme hautement « similaires ») ? Quel est le niveau de similarité qui nous paraît justifier une élimination et peut-on appliquer ce « barème » à l'ensemble du fonds ?

Pour les images similaires peut-être plus qu'ailleurs dans ce Mémoire de Master apparaît la délicate interaction homme-machine dans l'évaluation et le tri archivistiques au niveau *macro* et (semi-)automatisé de fonds d'archives privées : parce que les éléments sont tous uniques, semblables certes, mais différents ; parce que le logiciel ne fonctionne que selon les paramètres qu'on lui a imposés et qu'on a choisis ; parce que l'outil (actuel) ne peut pas intégrer tous les éléments *para*-informatiques qui peuvent influencer le choix de conserver telle ou telle image (le plan de classement initial, la structure de l'arborescence, mais aussi l'histoire de l'acquisition du Fonds, le parcours professionnel et personnel du créateur des documents, les parti-pris conscients et inconscients de l'archiviste en charge du tri, la politique de l'institution, etc.) ; enfin, et surtout, parce que l'outil se montre extraordinairement efficace pour l'identification des éléments, mais que la décision finale nous revient et nous incombe.

Tableau 29 : Comparaison de données, tâches

Tâche	Sous-tâches		Méthodes et instruments	Points d'analyse	Outils
Comparaisons de données	Dossiers		<ul style="list-style-type: none"> <li>* Visualisation de l'arborescence</li> <li>* Checksums</li> <li>* Recherche par nom de dossiers</li> <li>* Comparaison de répertoires</li> </ul>	<ul style="list-style-type: none"> <li>* Quantité d'éléments identiques / distincts</li> <li>* Type d'éléments identiques / distincts</li> <li>* Date de modification des éléments distincts</li> <li>* Plan de classification initial</li> </ul>	<ul style="list-style-type: none"> <li>* Archifiltre (visualisation)</li> <li>* TreeSize (redondances strictes et comparaison de métadonnées)</li> <li>* Beyond Compare (comparaison de répertoires et de dossiers)</li> </ul>
	Fichiers	Chemin unique	<ul style="list-style-type: none"> <li>* Recherche par métadonnées identiques</li> <li>* Algorithmes de comparaisons</li> <li>* Comparaison fine (bloc / ligne)</li> </ul>	<ul style="list-style-type: none"> <li>* Date de modification des éléments distincts</li> <li>* Seuil de tolérance (pourcentage de similarité)</li> <li>* Nomenclature des éléments</li> <li>* Format des éléments</li> </ul>	<ul style="list-style-type: none"> <li>* TreeSize (comparaison de métadonnées)</li> <li>* AllDup / AntiDupl (comparaison du contenu - images)</li> <li>* Beyond Compare (comparaison du contenu - textes)</li> </ul>
		Chemins multiples		<ul style="list-style-type: none"> <li>* Date de modification des éléments distincts</li> <li>* Seuil de tolérance (pourcentage de similarité)</li> <li>* Nomenclature des éléments</li> <li>* Format des éléments</li> <li>* Plan de classification initial</li> </ul>	



## 5. Conclusion

Au terme de cette recherche, nous voudrions établir une synthèse des principaux résultats et proposer quelques réflexions d'ordre plus général. Tout d'abord, grâce aux tests de onze logiciels, actifs dans des domaines différents et développés dans des contextes et des environnements distincts, nous sommes en mesure de recommander l'utilisation des outils suivants à la Cinémathèque suisse :

- *Karen's Directory Printer* et *TreeSize* pour l'extraction de métadonnées et l'établissement d'un récolement ;
- *Archifiltre* et *TreeSize* pour l'analyse de l'arborescence (*Droid* peut être utile pour l'identification précise des formats de fichiers) ;
- *AllDup* et *TreeSize* pour la recherche de redondances strictes ;
- *Beyond Compare* pour la comparaison de répertoires, *TreeSize* pour la recherche de métadonnées, et *AntiDupl* pour le traitement des images similaires.

Ces outils permettront à la Cinémathèque de mettre en œuvre les cinq tâches que nous avons proposées dans la méthodologie de tri archivistique portant sur les supports de données complexes : l'extraction, l'analyse de l'arborescence, le traitement des dossiers vides, le traitement des redondances strictes et la comparaison de données. Cependant, comme nous l'avons vu tout au long de cette étude, les outils ne répondent pas à toutes les attentes et à tous les besoins d'une institution patrimoniale en termes de traitement : d'une part, il est souvent nécessaire de cumuler l'utilisation de plusieurs outils ; d'autre part, ces outils ne proposent pas les analyses détaillées qui seules permettront de prendre des décisions éclairées. Ils doivent donc être accompagnés d'analyses statistiques, de visualisations graphiques, de repérages manuels, de tests sur des échantillons seulement, etc. Grâce à la liste des micro-fonctionnalités pondérées par pertinence pour le tri archivistique, il sera possible d'établir le cahier des charges d'un outil « idéal » – quelques-unes des analyses proposées, comme la base de données relationnelle, ou des statistiques plus détaillées, pourraient d'ailleurs y être intégrées.

Si les analyses proposées par les logiciels ne sont pas encore à la hauteur des attentes, quantité d'outils peuvent en revanche être utilisés sous forme d'aide à l'identification et pour une première prise de connaissance : ils permettent effectivement de découvrir de manière relativement aisée des contenus enregistrés sur des supports de données complexes inaccessibles sans l'intermédiaire d'un ordinateur ou d'un programme adapté. Pour ce qui est des traitements proprement dits, on a pu constater que plusieurs fonctionnalités proposées n'étaient tout simplement pas applicables dans le domaine archivistique, car elles vont à l'encontre des principes qui sous-tendent la discipline et sont contraires à la déontologie archivistique – on pense à la déduplication « en masse », ou à la fusion des dossiers et fichiers texte similaires proposée par exemple par *Beyond Compare* et *WinMerge*, qui font fi des critères de qualité propres aux documents d'archives comme étant des traces authentiques, intègres et fiables des activités du producteur.

Si plusieurs outils différents doivent être utilisés en parallèle, il en va de même pour ce qui est de la méthodologie. En effet, alors que certaines étapes doivent obligatoirement être réalisées de manière linéaire et sont prioritaires (comme l'élimination des fichiers système, la décompression des dossiers conteneurs, ou encore l'extraction initiale des métadonnées),

d'autres processus doivent être menés de manière itérative : si l'analyse de l'arborescence permet par exemple de décider de la conservation de telle ou telle instance d'un élément dont il existe plusieurs exemplaires, la recherche de redondances strictes fait également apparaître une organisation interne au fonds d'archives qui n'avait peut-être pas été identifiée lors de l'étude générale de l'arborescence. Des allers-retours constants sont donc nécessaires, entre les différents outils, mais également entre les étapes de la méthodologie et entre les approches *macro-* et *microscopiques* : l'identification des formats de fichiers, l'étude de la composition des dossiers, l'analyse de la profondeur de l'arborescence, la vérification d'hypothèses sur un échantillon de données, etc. Si toutes les étapes doivent *in fine* être réalisées, les approches peuvent différer d'un moment à l'autre de l'évaluation, d'un répertoire à l'autre, d'un support de données à l'autre et, bien sûr, d'un fonds d'archives à l'autre : l'analyse de l'arborescence sera probablement moins détaillée pour des clés USB que pour des disques durs externes, la comparaison d'images se justifiera suivant le nombre de clichés conservés, la base de données relationnelle ne servira que dans les cas où les chemins contenant sont très divergents et qu'il n'est pas possible d'identifier visuellement des *pattern* dans l'affichage des résultats proposés par les logiciels, etc.

Ce travail a enfin permis de montrer toute la complexité de l'évaluation et du tri archivistiques. D'une part, on a pu constater que même si certaines tâches paraissent relativement triviales et faciles à réaliser (comme le traitement des dossiers vides) ou qu'elles semblent pouvoir être entièrement couvertes par une fonctionnalité informatique (comme le traitement des redondances), en réalité, il n'en est rien : l'évaluation d'un fonds d'archives numériques est très intimement liée avec les autres étapes ou fonctions archivistiques que sont la classification et la description, et il est nécessaire d'adopter une approche holistique, qui prenne en compte l'entier du fonds et son organisation interne avant de pouvoir commencer le traitement proprement dit. Dans le cas des redondances par exemple, on ne peut pas commencer leur traitement concret (conservation ou élimination) sans avoir pris une décision à propos du plan de classification : gardera-t-on strictement celui du producteur des documents, ou établira-t-on une nouvelle classification ? Enfin, le deuxième enseignement, qui est aussi une limite de ce travail et une réserve quant à notre approche : avec le numérique plus qu'ailleurs peut-être, il faut se méfier d'une forme d'illusion « analytique », « quantitative » ou « statistique » dans la gestion des archives numériques (et plus encore en ce qui concerne les archives privées). Si l'on peut considérer qu'il n'y a pas de « boîte noire » avec les archives numériques, puisqu'une extraction de métadonnées permet, en quelques clics, d'avoir une liste de tous les dossiers et fichiers que contient un fonds, cette connaissance première, non hiérarchisée, n'est que le « préalable » à la prise de décision réelle. On peut certes multiplier les analyses statistiques grâce à l'immense masse d'informations dont nous disposons à propos du contenu des supports de données complexes, mais, comme nous l'évoquions pour le traitement des images similaires, aucun logiciel, aucune visualisation, ou aucun tableau statistique ne décidera pour nous du sort final à réserver aux documents. Ce Mémoire de Master traite donc bel et bien du « tri » archivistique, pensé comme l'*opération* qui consiste à séparer les documents qui doivent être conservés, de ceux qui seront détruits, mais l'acte de *juger* de la valeur de ces documents, de prendre la décision, est une toute autre affaire...

## Bibliographie

AIMS WORK GROUP, 2012. *AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship* [en ligne]. [Consulté le 15 mars 2022]. Disponible à l'adresse: <https://dcs.library.virginia.edu/aims/white-paper/>

ANTIDUPL, 2020a. Options Window. *AntiDupl.NET Description*. [en ligne]. 2020. [Consulté le 13 août 2022]. Disponible à l'adresse: <https://ermig1979.github.io/AntiDupl/data/help/english/index.html?page=options.html>

ANTIDUPL, 2020b. Frequently Asked Questions. *AntiDupl.NET Description*. [en ligne]. 2020. [Consulté le 13 août 2022]. Disponible à l'adresse: <https://ermig1979.github.io/AntiDupl/data/help/english/index.html?page=faq.html>

ANTIDUPL, 2020c. Table of Results. *AntiDupl.NET Description*. [en ligne]. 2020. [Consulté le 13 août 2022]. Disponible à l'adresse: <https://ermig1979.github.io/AntiDupl/data/help/english/index.html?page=table.html>

BELOVARI, Susanne, 2017. Expedited digital appraisal for regular archivists: an MPLP-type approach. *Journal of Archival Organization* [en ligne]. 3 avril 2017. Vol. 14, no. 1-2, pp. 55-77. [Consulté le 15 mars 2022]. DOI 10.1080/15332748.2018.1503014. Disponible à l'adresse : <https://doi.org/10.1080/15332748.2018.1503014>

BENEDETTI, Julien, 2021. Carte des outils utiles pour l'archivage électronique. *Association des archivistes français*. [en ligne]. 4 novembre 2021. [Consulté le 11 juillet 2022]. Disponible à l'adresse: <https://www.archivistes.org/Carte-des-outils-utiles-pour-l-archivage-electronique>

BIASI, Pierre-Marc, 2013. Les sentiers perdus de la création. In: *Literaturarchiv - literarisches Archiv: zur Poetik literarischer Archive = Archives littéraires et poétiques d'archives : écrivains et institutions en dialogue*. Göttingen: Wallstein. pp. 35-49. Beide Seiten Bd. 3. ISBN 978-3-0340-1156-3.

CHASSANOFF, Alexandra et POST, Colin, 2020. *OSSArcFlow. Guide to Documenting Born-Digital Archival Workflows* [en ligne]. Atlanta: Educopia Institute. [Consulté le 19 mai 2022]. Disponible à l'adresse: <https://educopia.org/ossarcflow-guide/>

CINÉMATHEQUE SUISSE, [sans date]. Mission. *cinematheque.ch*. [en ligne]. [Consulté le 29 juillet 2022 a]. Disponible à l'adresse: <https://www.cinematheque.ch/f/cinematheque-en-bref/organisation/conseil-de-fondation/mission/>

CINÉMATHEQUE SUISSE, [sans date]. Cinémathèque en bref. *cinematheque.ch*. [en ligne]. [Consulté le 11 août 2022 b]. Disponible à l'adresse: <https://www.cinematheque.ch/f/cinematheque-en-bref/>

CINÉMATHEQUE SUISSE, 2015a. La politique de collection. *cinematheque.ch*. [en ligne]. 25 mars 2015. [Consulté le 27 juillet 2022]. Disponible à l'adresse: <https://www.cinematheque.ch/f/espace-institutionnel/>

CINÉMATHEQUE SUISSE, 2015b. Le numérique à la Cinémathèque suisse. Synthèse au 1er septembre 2015. *cinematheque.ch*. [en ligne]. 1 septembre 2015. [Consulté le 11 août 2022]. Disponible à l'adresse: [https://www.cinematheque.ch/fileadmin/user\\_upload/Espace-institutionnel/CS\\_150901\\_synthesenumerique.pdf](https://www.cinematheque.ch/fileadmin/user_upload/Espace-institutionnel/CS_150901_synthesenumerique.pdf)

CINÉMATHEQUE SUISSE, 2020a. Rapport d'activités 2020. *cinematheque.ch*. [en ligne]. 2020. [Consulté le 27 juillet 2022]. Disponible à l'adresse: <https://www.cinematheque.ch/f/espace-institutionnel/>

CINÉMATHEQUE SUISSE, 2020b. Organigramme. *cinematheque.ch*. [en ligne]. 2020. [Consulté le 29 juillet 2022]. Disponible à l'adresse: <https://www.cinematheque.ch/f/cinematheque-en-bref/organisation/organigramme/>

CINÉMATHEQUE SUISSE, 2021. Rapport d'activités 2021. *cinematheque.ch*. [en ligne]. 2021. [Consulté le 27 juillet 2022]. Disponible à l'adresse: <https://www.cinematheque.ch/f/espace-institutionnel/>

CINÉMATHEQUE SUISSE, 2022a. Modalités des dépôts numériques à la Cinémathèque suisse. *cinematheque.ch*. [en ligne]. mai 2022. [Consulté le 27 juillet 2022]. Disponible à l'adresse: <https://www.cinematheque.ch/f/espace-institutionnel/>

CINÉMATHEQUE SUISSE, 2022b. Plateforme logicielle. Présentation générale. Lausanne. 24 mai 2022. Document interne [Fichier Power Point]

COLINE1, 2019. La boîte à outils numériques de l'archiviste. *Modernisation et archives*. [en ligne]. 11 décembre 2019. [Consulté le 12 avril 2022]. Disponible à l'adresse: <https://siaf.hypotheses.org/1059>

COMITÉ CONSULTATIF POUR LES SYSTÈMES DE DONNÉES SPATIALES (CCSDS), 2017. *Modèle de référence pour un Système ouvert d'archivage d'information (OAIS)*. [en ligne]. Washington D.C.: CCSDS Secretariat. [Consulté le 11 août 2022]. Disponible à l'adresse: <https://public.ccsds.org/Pubs/650x0m2%28F%29.pdf>

COMMUNITY OWNED DIGITAL PRESERVATION TOOL REGISTRY (COPTR), 2021. Tools Grid. *coptr.digipres.org*. [en ligne]. 27 octobre 2021. [Consulté le 20 avril 2022]. Disponible à l'adresse: [https://coptr.digipres.org/index.php/Tools\\_Grid](https://coptr.digipres.org/index.php/Tools_Grid)

CONTENT BLOCKCHAIN, 2019. Testing different image hash functions. *Content blockchain*. [en ligne]. 2019. [Consulté le 10 août 2022]. Disponible à l'adresse: <https://content-blockchain.org/research/testing-different-image-hash-functions/>

COUTAZ, Gilbert, 2016. La croissance et la maîtrise des masses documentaires. *arbido*. [en ligne]. 2016. No. 3. [Consulté le 19 mai 2022]. Disponible à l'adresse: <https://arbido.ch/fr/edition-article/2016/d%C3%A9truire-pour-conserver/la-croissance-et-la-ma%C3%AEtrise-des-masses-documentaires>

COUTURE, Carol, 1999. *Les fonctions de l'archivistique contemporaine*. Sainte-Foy, Québec: Presses de l'Université du Québec. Gestion de l'information. ISBN 2-7605-0941-9.

DI LENARDO, Isabella, SEGUIN, Benoît Laurent Auguste et KAPLAN, Frédéric, 2016. Visual Patterns Discovery in Large Databases of Paintings. In: *Digital Humanities 2016*. [en ligne]. Krakow. juillet 2016. [Consulté le 13 août 2022]. Disponible à l'adresse: <https://infoscience.epfl.ch/record/220638?ln=fr> [Preprint]

DUCHARME, Daniel, 2001. L'identification de critères d'évaluation pour les archives informatiques. *Archives* [en ligne]. 2001. Vol. 32, no. 2, pp. 17-32. [Consulté le 15 août 2022]. Disponible à l'adresse : [https://www.archivistes.qc.ca/revuearchives/vol32\\_2/32-2-ducharme.pdf](https://www.archivistes.qc.ca/revuearchives/vol32_2/32-2-ducharme.pdf)

FABRIQUE NUMÉRIQUE DES MINISTÈRES SOCIAUX, 2021. Wiki Archifiltre. *GitHub*. [en ligne]. 2021. [Consulté le 29 juillet 2022]. Disponible à l'adresse: <https://github.com/SocialGouv/archifiltre-docs>

FORSTROM, Michael, 2009. Managing Electronic Records in Manuscript Collections: A Case Study from the Beinecke Rare Book and Manuscript Library. *The American Archivist*. [en ligne].



2009. Vol. 72, no. 2, pp. 460-477. DOI 10.17723/aarc.72.2.b82533tvr7713471. [Consulté le 18 mars 2022]. Disponible à l'adresse : <https://doi.org/10.17723/aarc.72.2.b82533tvr7713471>

FORTIN, Marie Fabienne, 2016. *Fondements et étapes du processus de recherche: méthodes quantitatives et qualitatives*. 3e édition. Montréal: Chenelière Education. ISBN 978-2-7650-5006-3.

FRANÇOIS, Robin, 2021. Flux de traitements des lots de données au Département Non-Film. Penthaz. 30 août 2021. Document interne [Fichier graphique]

FRANÇOIS, Robin et ROCHAT, Rebecca, 2022. Digital Preservation Pipeline for Data Storage Media At The Cinemathèque Suisse. Imaging and extracting data and metadata from Special Collections media. In: *iPres 2022: The 18th International Conference on Digital Preservation*. Glasgow. septembre 2022. [En cours de publication].

GAUDINAT, Arnaud, 2016. Le plaisir de tout conserver sans modération: une question de taille ? *arbido*. [en ligne]. N° 3 2016. [Consulté le 15 mars 2022]. Disponible à l'adresse: <https://arbido.ch/fr/edition-article/2016/d%C3%A9truire-pour-conserver/le-plaisir-de-tout-conserver-sans-mod%C3%A9ration-une-question-de-taillehttps://arbido.ch/fr/>

GRAND DICTIONNAIRE TERMINOLOGIQUE, 2012. Arborescence. *Grand dictionnaire terminologique (GDT)*. [en ligne]. 2012. [Consulté le 13 août 2022]. Disponible à l'adresse: [https://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id\\_Fiche=8369503](https://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=8369503)

GRAND DICTIONNAIRE TERMINOLOGIQUE (GDT), 2012. Fonctionnalité. *Grand dictionnaire terminologique (GDT)*. [en ligne]. 2012. [Consulté le 13 août 2022]. Disponible à l'adresse: [https://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id\\_Fiche=26559312](https://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=26559312)

GREENE, Mark et MEISSNER, Dennis, 2005. More Product, Less Process: Revamping Traditional Archival Processing. *The American Archivist* [en ligne]. 1 septembre 2005. Vol. 68, no. 2, pp. 208-263. [Consulté le 14 août 2022]. DOI 10.17723/aarc.68.2.c741823776k65863. Disponible à l'adresse : <https://doi.org/10.17723/aarc.68.2.c741823776k65863>

INTERPARES, 2018. Metadata. *InterPARES Trust AI - Artificial Intelligence - Terminology Database*. [en ligne]. 2018. [Consulté le 12 août 2022]. Disponible à l'adresse: <https://interparestrustai.org/terminology/term/metadata/en>

JAM SOFTWARE (JOACHIM MARDER), 2022. TreeSize. *jam-software.com*. [en ligne]. 2022. [Consulté le 31 juillet 2022]. Disponible à l'adresse: <https://manuals.jam-software.com/treesize/EN/PDF/TreeSize.pdf>

KARENWARE, 2022. Karen's Power Tools. *Karen's Power Tools*. [en ligne]. 2022. [Consulté le 10 août 2022]. Disponible à l'adresse: <https://www.karenware.com/>

KIM, Sarah, DONG, Lorraine A. et DURDEN, Megan, 2006. Automated Batch Archival Processing: Preserving Arnold Wesker's Digital Manuscripts. *Archival Issues* [en ligne]. 2006. Vol. 30, no. 2, pp. 91-106. [Consulté le 27 mars 2022]. Disponible à l'adresse : <https://www.jstor.org/stable/41102125>

KOST-CECO, 2022a. Analyse. *KOST Wiki*. [en ligne]. 19 avril 2022. [Consulté le 12 août 2022]. Disponible à l'adresse: <https://www.kost-ceco.ch/kostwiki/doku.php?id=analyse>

KOST-CECO, 2022b. Tools. *KOST Wiki*. [en ligne]. 13 mai 2022. [Consulté le 21 juin 2022]. Disponible à l'adresse: <https://kost-ceco.ch/kostwiki/doku.php?id=tools>

LANGDON, John, 2016. Describing the digital: the archival cataloguing of born-digital personal papers. *Archives and Records* [en ligne]. 2 janvier 2016. Vol. 37, no. 1, pp. 37-52. [Consulté le 14 août 2022] DOI 10.1080/23257962.2016.1139494. Disponible à l'adresse : <https://doi.org/10.1080/23257962.2016.1139494>

Loi fédérale sur la Bibliothèque nationale suisse (LBNS ; 423.21), 1992. *Fedlex. La plateforme de publication du droit fédéral*. [en ligne]. 18 décembre 1992. Mise à jour le 1<sup>er</sup> février 2021. [Consulté le 25 juillet 2022]. Disponible à l'adresse: [https://www.fedlex.admin.ch/eli/cc/1993/1773\\_1773\\_1773/fr](https://www.fedlex.admin.ch/eli/cc/1993/1773_1773_1773/fr)

Loi fédérale sur la culture et la production cinématographiques (LCin ; 443.1), 2001. *Fedlex. La plateforme de publication du droit fédéral*. [en ligne]. 14 décembre 2001. Mise à jour le 1<sup>er</sup> janvier 2022. [Consulté le 25 juillet 2022]. Disponible à l'adresse: <https://www.fedlex.admin.ch/eli/cc/2002/283/fr>

MAIRE, Frédéric, 2019. Décès d'Ana Simon. *Cinemathèque suisse*. [en ligne]. 21 janvier 2019. [Consulté le 11 août 2022]. Disponible à l'adresse: <https://www.cinematheque.ch/i/actualites/article/deces-dana-simon/>

MAIRE, Frédéric, 2020. Francis Reusser nous a quittés. *Cinemathèque suisse*. [en ligne]. 4 octobre 2020. [Consulté le 11 août 2022]. Disponible à l'adresse: <https://www.cinematheque.ch/i/actualites/article/francis-reusser-nous-a-quittes/>

MAKHLOUF SHABOU, Basma, 2011. Étude sur la définition et la mesure des qualités des archives définitives issues d'une évaluation. *Archives* [en ligne]. 2012 2011. Vol. 43, no. 2, pp. 39-70. [Consulté le 17 mars 2022]. Disponible à l'adresse : [https://www.archivistes.qc.ca/revuearchives/vol43\\_2/43\\_2\\_makhlouf-shabou.pdf](https://www.archivistes.qc.ca/revuearchives/vol43_2/43_2_makhlouf-shabou.pdf)

MAKHLOUF SHABOU, Basma, 2021. *Introduction à l'OAIS* [Présentation Powerpoint]. Genève. 11 novembre 2021. Support de cours : M7c Gouvernance des données, Haute École de Gestion, 2021.

MEISTER, Sam et CHASSANOFF, Alexandra, 2014. Integrating Digital Forensics Techniques into Curatorial Tasks: A Case Study. *International Journal of Digital Curation* [en ligne]. 30 octobre 2014. Vol. 9, no. 2, pp. 6-16. [Consulté le 30 mai 2022]. DOI 10.2218/ijdc.v9i2.325. Disponible à l'adresse : <http://www.ijdc.net/article/view/9.2.6>

MIEGEL, Annekathrin, SCHIEBER, Sigrid et SCHMIDT, Christoph, 2017. Vom richtigen Umgang mit kreativen digitalen Ablagen. In: *Kreative digitale Ablagen und die Archive: Ergebnisse eines Workshops des KLA-Ausschusses Digitale Archive am 22./23.11.2016 in der Generaldirektion der Staatlichen Archive Bayerns*. München: Generaldirektion der staatlichen Archive Bayerns. 2017. pp. 7-16. Sonderveröffentlichungen der Staatlichen Archive Bayerns. ISBN 978-3-938831-81-6.

MORISOD, Pascal, 2018. Des archives, des machines et des hommes, un heureux ménage à trois ? *arbido*. [en ligne]. 2018. [Consulté le 15 août 2022]. Disponible à l'adresse: <https://arbido.ch/fr/edition-article/2018/automatisation-versprechen-oder-drohung-des-archives-des-machines-et-des-hommes-un-heureux-m%C3%A9nage-%C3%A0-trois>

MTSD, 2022. *Find Similar Pictures. AllDup*. 2022. [Aide contextuelle du logiciel]

NEESER, Caroline, 2022. Fonds Francis Reusser - Inventaire des archives papier de la Cinémathèque suisse. *Caspar. Inventaire des archives papier*. [en ligne]. février 2022. [Consulté le 11 août 2022]. Disponible à l'adresse: <https://caspar.cinematheque.ch/reusser>

NESTOR - DEUTSCHE NATIONALBIBLIOTHEK, 2022. Werkzeugübersicht. *Nestor-Wiki*. [en ligne]. 29 mars 2022. [Consulté le 30 juin 2022]. Disponible à l'adresse: <https://wiki.dnb.de/pages/viewpage.action?pageId=134715087>

OESTREICHER, Cheryl, 2013. Personal Papers and MPLP: Strategies and Techniques. *Archivaria* [en ligne]. 2013. pp. 93-110. [Consulté le 16 mars 2022]. Disponible à l'adresse : <https://archivaria.ca/index.php/archivaria/article/view/13460>

Ordonnance sur la Bibliothèque nationale suisse (OBNS ; 432.211), 1998. *Fedlex. La plateforme de publication du droit fédéral*. [en ligne]. 14 janvier 1998. Mise à jour le 1<sup>er</sup> janvier 2021. [Consulté le 25 juillet 2022]. Disponible à l'adresse: [https://www.fedlex.admin.ch/eli/cc/1998/204\\_204\\_204/fr](https://www.fedlex.admin.ch/eli/cc/1998/204_204_204/fr)

PARADIGM PROJECT, 2007. *Workbook on Digital Private Papers* [en ligne]. Oxford: University of Oxford and University of Manchester. [Consulté le 9 avril 2022]. Disponible à l'adresse: <https://ora.ox.ac.uk/objects/uuid:116a4658-deff-4b06-81c5-c9c2071bc6d0>

PCMAG, 2022. file manager. *PCMag Encyclopedia*. [en ligne]. 2022. [Consulté le 13 août 2022]. Disponible à l'adresse: <https://www.pcmag.com/encyclopedia/term/file-manager>

PORTAIL INTERNATIONAL ARCHIVISTIQUE FRANCOPHONE (PIAF), 2015a. Tri. *Glossaire*. [en ligne]. 2015. [Consulté le 15 août 2022]. Disponible à l'adresse: [https://www.piaf-archives.org/sites/default/files/bulk\\_media/glossaire/co/Module\\_glossaire\\_16.html](https://www.piaf-archives.org/sites/default/files/bulk_media/glossaire/co/Module_glossaire_16.html)

PORTAIL INTERNATIONAL ARCHIVISTIQUE FRANCOPHONE (PIAF), 2015b. Récolement. *Glossaire*. [en ligne]. 2015. [Consulté le 13 août 2022]. Disponible à l'adresse: [https://www.piaf-archives.org/sites/default/files/bulk\\_media/glossaire/co/Module\\_glossaire\\_14.html#footnotesN153](https://www.piaf-archives.org/sites/default/files/bulk_media/glossaire/co/Module_glossaire_14.html#footnotesN153)

PORTAIL INTERNATIONAL ARCHIVISTIQUE FRANCOPHONE (PIAF), 2015c. Empreinte. *Glossaire*. [en ligne]. 2015. [Consulté le 13 août 2022]. Disponible à l'adresse: [https://www.piaf-archives.org/sites/default/files/bulk\\_media/glossaire/co/Module\\_glossaire\\_5.html#footnotesN154](https://www.piaf-archives.org/sites/default/files/bulk_media/glossaire/co/Module_glossaire_5.html#footnotesN154)

POST, Colin, CHASSANOFF, Alexandra, LEE, Christopher, RABKIN, Andrew, ZHANG, Yinglong, SKINNER, Katherine et MEISTER, Sam, 2019. Digital Curation at Work: Modeling Workflows for Digital Archival Materials. In: *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. juin 2019. pp. 39-48. [Consulté le 12 mai 2022]. DOI 10.1109/JCDL.2019.00016. Disponible à l'adresse : <https://ieeexplore.ieee.org/document/8791228>

ROCHAT, Rebecca, 2021. Flux numérique, Archivage / INGEST, Non-Film. Penthaz. 16 mars 2021. Document interne [Fichier graphique]

ROTH-LOCHNER, Barbara et GISLER, Johanna, 2007. Accroissement et collecte: les archives sur le marché patrimonial. In: COUTAZ, Gilbert (éd.), *Archivpraxis in der Schweiz = Pratiques archivistiques en Suisse*. Baden: Hier und Jetzt. pp. 303-318.

ROY, Simon N., 2010. L'étude de cas. In: *Recherche sociale : de la problématique à la collecte des données*. [en ligne]. Québec: Presses de l'Université du Québec. pp. 199-225. [Consulté le 11 août 2022]. ISBN 978-2-7605-1600-7. Disponible à l'adresse: <https://hesge.scholarvox.com/catalog/book/docid/88801643>

SCHNEIDER, J., ADAMS, C., DEBAUCHE, S., ECHOLS, R., MCKEAN, C., MORAN, J. et WAUGH, D., 2019. Appraising, processing, and providing access to email in contemporary literary archives. *Archives and Manuscripts* [en ligne]. 2 septembre 2019. Vol. 47, no. 3, pp. 305-326. [Consulté le 19 avril 2022]. DOI 10.1080/01576895.2019.1622138. Disponible à l'adresse : <https://doi.org/10.1080/01576895.2019.1622138>

SCOOTER SOFTWARE, 2022. *Paramètres divers de Comparaison de Dossiers. Beyond Compare* 2022. [Aide contextuelle du logiciel]

SEGUIN, Benoit, STRIOLO, Carlotta, DILENARDO, Isabella et KAPLAN, Frederic, 2016. Visual Link Retrieval in a Database of Paintings. In: HUA, Gang et JÉGOU, Hervé (éd.), *Computer Vision – ECCV 2016 Workshops* [en ligne]. Cham: Springer International Publishing. 2016. pp. 753-767. Lecture Notes in Computer Science. [Consulté le 13 août 2022]. ISBN 978-3-319-46604-0. DOI 10.1007/978-3-319-46604-0\_52. Disponible à l'adresse : [https://link.springer.com/chapter/10.1007/978-3-319-46604-0\\_52](https://link.springer.com/chapter/10.1007/978-3-319-46604-0_52)

SHEIN, Cyndi, 2014. From Accession to Access: A Born-Digital Materials Case Study. *Journal of Western Archives*. [en ligne]. 2014. Vol. 5, no. 1. [Consulté le 14 août 2022]. DOI <https://doi.org/10.26077/b3e2-d205>. Disponible à l'adresse : <https://digitalcommons.usu.edu/westernarchives/vol5/iss1/1>

SLOYAN, Victoria, 2016. Born-digital archives at the Wellcome Library: appraisal and sensitivity review of two hard drives. *Archives and Records* [en ligne]. 2 janvier 2016. Vol. 37, no. 1, pp. 20-36. [Consulté le 31 mars 2022]. DOI 10.1080/23257962.2016.1144504. Disponible à l'adresse : <https://doi.org/10.1080/23257962.2016.1144504>

SOCIETY OF AMERICAN ARCHIVISTS, 2005. Checksum. *The Dictionary of Archives Terminology*. [en ligne]. 2005. [Consulté le 13 août 2022]. Disponible à l'adresse: <https://dictionary.archivists.org/entry/checksum.html>

THE NATIONAL ARCHIVES, [sans date]. Detecting duplicate files. *Droid 6 Help*. [Aide contextuelle du logiciel]

TOURN, Christine, 2019. Papiers Ana Simon - Inventaire des archives papier de la Cinémathèque suisse. *Caspar. Inventaire des archives papier*. [en ligne]. octobre 2019. [Consulté le 11 août 2022]. Disponible à l'adresse: <https://caspar.cinematheque.ch/ana-simon>

TRACE, Ciaran B., 2021. Archival infrastructure and the information backlog. *Archival Science*. [en ligne] 21 juillet 2021. Vol. 22, no. 1, pp. 75-93. [Consulté le 22 avril 2022]. DOI 10.1007/s10502-021-09368-x. Disponible à l'adresse : <https://doi.org/10.1007/s10502-021-09368-x>

VINH-DOYLE, William P., 2017. Appraising email (using digital forensics): techniques and challenges. *Archives and Manuscripts* [en ligne]. 2 janvier 2017. Vol. 45, no. 1, pp. 18-30. [Consulté le 4 avril 2022]. DOI 10.1080/01576895.2016.1270838. Disponible à l'adresse : <https://doi.org/10.1080/01576895.2016.1270838>

WANG, Zhou, BOVIK, Alan Conrad, SHEIKH, Hamid Rahim et SIMONCELLI, Eero P., 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*. avril 2004. Vol. 13, no. 4, pp. 600-612. [Consulté le 7 août 2022]. DOI 10.1109/TIP.2003.819861. Disponible à l'adresse : <http://ieeexplore.ieee.org/document/1284395/>

WILSEY, Laura, SKIRVIN, Rebecca, CHAN, Peter et EDWARDS, Glynn, 2013. Capturing and Processing Born-Digital Files in the STOP AIDS Project Records: A Case Study. *Journal of Western Archives*. [en ligne]. 26 avril 2013. Vol. 4, no. 1. [Consulté le 10 mai 2022].

DOI <https://doi.org/10.26077/43de-194f>. Disponible à l'adresse :  
<https://digitalcommons.usu.edu/westernarchives/vol4/iss1/1>

WINMERGE, [sans date]. Comparing folders. *WinMerge 2.16 Help*. [en ligne].  
[Consulté le 13 août 2022]. Disponible à l'adresse:  
[https://manual.winmerge.org/en/Quick\\_start.html#id591190](https://manual.winmerge.org/en/Quick_start.html#id591190)

YOUNG, Julia Marks et BOLES, Frank, 1991. *Archival appraisal*. New York: Neal-Schuman Publishers. ISBN 978-1-55570-064-5.