Prise en main automatisée du vrac numérique audiovisuel d'un service communication

Sara DA SILVA SANTOS
s.dasilvasantos@protonmail.com
https://orcid.org/0009-0009-5205-6720
Archiviste à la galerie De Jonckheere, Genève

Fleur HEINIGER
https://orcid.org/0009-0000-8056-8402
fleurhheiniger@gmail.com
Auxiliaire à la bibliothèque Ernst & Lucie Schmidheiny (Sciences II), Université de Genève

Résumé

La prise en main d'un vrac numérique est complexe en raison de l'ampleur et la diversité de ses données non structurées.

Notre recherche consiste à étudier à travers un cas pratique, le contexte et les défis du vrac numérique du service communication d'une institution publique, ainsi que les risques encourus. Pour cela, nous avons utilisé les outils d'automatisation open source Archifiltre et DROID. Le but étant d'établir la manière de désengorger un vrac numérique et de prévenir sa réapparition.

Nous avons étudié l'état du disque, l'arborescence, sa nomenclature, diagnostiqué les données et la répartition des formats.

Nos résultats montrent que le vrac est constitué de formats audiovisuels majoritairement propriétaires, dont un grand nombre de rushs.

À l'issue de ce projet, nous avons identifié trois types d'axes d'amélioration : les pratiques du service, la gestion du volume des données ainsi que quelques critères utiles à la valorisation future des données.

Mots-clés

Vrac numérique, audiovisuel, automatisation, évaluation, valorisation, patrimonial

1. Introduction

La prise en main d'un vrac numérique est une étape complexe qui ne s'improvise pas. L'ampleur et la diversité des données non structurées qui le constituent représentent un challenge de taille pour les institutions (Texier, 2021b), tant et si bien que Gilbert Coutaz parle de « labyrinthe informationnel complexe » (Coutaz, 2016b). Les données contenues sont souvent de sources multiples et formats variés. Le plus souvent, elles ne sont pas intégrées dans un système de classement compréhensif, décrites avec métadonnées complètes et suivant des règles de nommage uniforme. Cet ensemble de caractéristiques façonne une collection de données numériques qui croît, sans véritable contrôle, à la façon d'un jardin laissé à l'abandon. Le vrac numérique peut atteindre des proportions qui deviennent difficiles à gérer a posteriori. Les institutions sont alors confrontées à une problématique d'envergure : la maîtrise de leurs données qui renferment un potentiel inexploité (Texier, 2022a). En effet, ces données pourraient aider à la prise de décision stratégique, favoriser l'innovation ou encore alimenter les archives historiques (Makhlouf Shabou et al., 2020 ; Makhlouf Shabou, 2023a).

La maîtrise du vrac numérique et de sa valeur informationnelle revêt par conséquent une importance stratégique capitale. Elle s'intègre à une compréhension complète de la production de l'institution, et est couplée à la sensibilisation des producteur·rice·s et des propriétaires à la présence d'actifs informationnels dans leurs fonds. Cette maîtrise passe par des fonctions archivistiques — telles la création ou l'évaluation — qui permettent de répondre à divers enjeux cruciaux. Ainsi, la création consiste à contrôler « la qualité, la validité, la crédibilité et la pérennité de l'information » dès la capture d'un document (Gagnon-Arguin, 2003, p. 81), une étape incontournable à la sensibilisation des productrices et producteurs pour une conception documentaire éclairée. L'évaluation quant à elle consiste à juger de la valeur d'un document analogique ou numérique afin de lui attribuer des durées de conservation — six mois, une année, dix ans, etc. — en fonction de la valeur établie. En résumé, l'évaluation se rapproche d'un « droit de vie ou de mort » sur les documents (Couture, 1996).

De plus, l'évaluation en tant que fonction archivistique possède de nombreux atouts dans un contexte institutionnel. Elle garantit le respect des normes et des réglementations, assurant la conformité des exigences en matière d'évaluation, de conservation et d'archivage des documents (International Standard Organisation, 2012, 2016, 2020, 2022a, 2022c, 2022b). Elle assure également la confidentialité et la sécurité des données en ne conservant que les documents nécessaires, réduisant ainsi les risques de violation de la vie privée et les fuites d'informations sensibles (L'Assemblée fédérale de la Confédération suisse, 2020 ; République et canton de Genève, 2000 a, 2021). En contrôlant et en triant efficacement les données, les institutions réduisent les coûts liés aux supports et aux structures de stockages. En adoptant une approche responsable du numérique, les institutions contribuent à réduire leur impact environnemental. Finalement, les collaboratrices et collaborateurs gagnent un temps précieux en évitant les recherches fastidieuses. En effet, il est estimé qu'un individu, dans le cadre de son travail, perdrait plus de sept heures par semaine à rechercher des fichiers particuliers (Texier, 2022).

Cependant, le contrôle de toutes les données produites lors des activités courantes d'une institution peut être difficile, notamment par manque de personnel, de temps ou de connaissances — techniques, technologiques, législatives ou archivistiques (Belovari, 2019, p. 55 et 57). De nombreuses institutions se retrouvent avec des masses de données difficiles

à naviguer et trop souvent inexploitables. L'automatisation peut apporter une réponse efficace et tangible au travail quotidien des archivistes, et optimiser la gouvernance de l'information de l'institution (Makhlouf Shabou et al., 2020).

Divers outils d'automatisation intégrant l'intelligence artificielle sont développés afin de soulager et simplifier certaines tâches répétitives de la gestion des données. Ils apportent de nouvelles perspectives en matière d'analyse, de traitement et d'évaluation des données (Makhlouf Shabou, 2023a) et représentent une réelle opportunité de soutien au traitement d'un vrac numérique. Cependant, tous ne sont pas aisés à prendre en main et peuvent nécessiter des connaissances techniques ou informatiques préalables, en plus de parfois présenter un coût non négligeable. De plus, comme l'a pointé un travail réalisé en 2022, il n'existe pas d'index exhaustif et mis à jour, recensant tous les outils d'automatisation existants, afin d'aider les spécialistes et les non-spécialistes à choisir ce qu'il leur faudrait en fonction de leurs besoins (Bavaud et al., 2022). Et malgré l'offre existante, aucun outil ne combine toutes les fonctions archivistiques propres aux données numériques. Autant de facteurs qui ralentissent la mise en place de ces outils automatisés dans la maîtrise des vracs numériques (Bavaud et al., 2022). Pourtant, ces outils représentent un potentiel de développement pour les institutions patrimoniales, régulièrement confrontées à des fonds audiovisuels d'origines diverses, qui nécessitent d'être évalués, avant l'intégration aux collections ou le versement aux archives (Bennani et al., 2022).

Cet article contribue à la recherche sur l'analyse automatisée d'un vrac numérique en se penchant sur la production de données audiovisuelles d'un service administratif, et plus particulièrement d'un département de communication et relations publiques. Ce service, présent dans toutes les institutions, produit de larges volumes de données et se trouve dans une zone grise concernant le traitement et la gestion de leurs données audiovisuelles. À l'inverse, les autres services administratifs — majoritairement producteurs de données textuelles ou de formats bureautiques — font l'objet de critères d'évaluation spécifiques connus comme l'ISO 30302:2022, ce dont manquent cruellement les services comme celui de la communication.

Afin de mieux comprendre les problèmes que rencontrent les services producteurs situés en zone grise, nous nous sommes penchés sur le cas pratique du service communication d'une institution publique (partie 1.5). À travers l'automatisation, la description des résultats et l'analyse de ce cas d'étude, nous souhaitons apporter des clés aux services similaires ayant les mêmes problématiques de vrac et d'évaluation.

1.1. Objectifs

Ce projet consiste en une recherche appliquée à travers un cas pratique, dans le but de comprendre le contexte de création de vrac numérique, les défis soulevés, les risques encourus par la situation actuelle de notre mandant et les opportunités possibles pour les données.

Nos objectifs :

- Effectuer une revue de la littérature
- Repérer des outils utiles pour l'analyse et l'évaluation du vrac
- Effectuer une analyse des besoins du mandant

- Extraire des rapports d'analyse automatisés des données stockées
- Proposer un cadre pratique pour une gestion documentaire dès les débuts du cycle de vie des documents

Ces objectifs visent à répondre aux questions de recherche listées au point suivant.

1.2. Question de recherche

Notre question de recherche consiste à déterminer les étapes de prise en main d'un vrac numérique contenant en grande majorité des données non structurées. Notre cas d'étude concernant le service de communication d'une institution publique, nous avons donc fait trois hypothèses :

- Les données sont majoritairement constituées de textes et de formats audiovisuels,
- Ils contiennent des formats, poids et versions hétérogènes
- Le service est peut-être confronté à une gestion documentaire lacunaire ou absente.

Il s'agit pour nous de répondre aux questions :

- Quelles sont les étapes pour la prise en main d'un vrac numérique ?
- Quelle est la plus-value de l'automatisation dans l'analyse et l'évaluation de grands volumes de documents ?
- Quels sont les outils open source pouvant assister les spécialistes et les services administratifs dans cette tâche ?
- Quelles sont les obligations spécifiques d'un service de communication, dont la mission est de promouvoir l'institution et ses activités, par rapport aux archives ?
- Quelles recommandations peuvent être faites pour prévenir la nouvelle formation de vrac numérique ?

1.3. Définition des concepts principaux

Pour répondre au mieux à ces questions, nous avons identifié des concepts clés que nous clarifions ci-après. En effet, la terminologie a parfois été complexe. Ceci s'explique par le fait que :

- L'état de la recherche sur la prise en main de vracs numériques est encore immature,
- La littérature est majoritairement anglophone.

Afin de faciliter la reproductibilité de notre revue de la littérature, nous avons indiqué les termes anglophones correspondants.

1.3.1. Valeur — Value (EN)

Degré pour lequel un ensemble de caractéristiques inhérentes à un objet répondent aux exigences[1]. Ces dernières correspondent à un besoin ou une attente exprimée[2] (Organisation Internationale de Normalisation, 2022a, p. 1).

[1] «3.1.3 quality degree to which a set of inherent characteristics of an object fulfils requirements (3.1.2)».
[2] « 3.1.2 requirement need or expectation that is stated, generally implied or obligatory ».

1.3.2. Actif informationnel — Information asset (EN)

Documents définis par une valeur économique ou stratégique aux yeux de son propriétaire[3] (International Standard Organisation, 2022b, p. 2; InterPARES Trust AI, 2023). Ils sont le plus souvent définis par leur valeur financière, qui atteste, démontre et appuie la propriété intellectuelle ou foncière du propriétaire (Makhlouf Shabou, 2023b). Leur importance stratégique et économique réside dans l'impact qu'ils ont sur la prise de décision, la compétitivité, l'innovation, la conformité réglementaire, la gestion des risques, et la capacité d'une organisation à s'adapter rapidement aux changements et à créer de la valeur ajoutée. Ils incluent toute donnée, tout dispositif ou tout composant qui soutient les activités liées à l'information, par exemple des brevets, logiciels, contrats, actes de propriété, mais aussi licences, logiciels, matériel informatique, serveurs, etc. (InterPARES Trust AI, 2023, déf. Wikipedia).

1.3.3. Vrac numérique — Data swamp (EN)[4]

Désigne des éléments disponibles dans un certain désordre ou « pêle-mêle » (lerobert.com, 2023). Étendu au domaine du numérique, le vrac se réfère à un ensemble de données non structurées, mal classées, mal nommées ou encore non identifiées, dans des formats multiples (Texier 2021 ; 2022 b). L'Association des archivistes français souligne les principaux problèmes que soulève le vrac : difficulté d'identification, impossibilité de lecture ou encore volumétrie conséquente. Le vrac numérique répond à un ou plusieurs critères (Association des archivistes français, 2018) :

- Peu ou pas de contexte de production
- Peu ou pas organisé
- Peu ou pas identifié
- Avec des documents de différents statuts ou versionnage[5]
- Contenus, formats et volumétries de fichiers hétérogènes
- Peu ou pas de métadonnées

Le vrac numérique survient le plus souvent dans des contextes de sauvetage d'un fonds comme lors du départ d'un·e membre du personnel ou la découverte d'un disque dur ou serveur oublié lors d'un déménagement de service.

1.3.4. Mégadonnées — Big Data (EN)

Données ne pouvant pas être approchées par un traitement manuel ou logiciel habituel, en raison de leur volume, leur variété et leur constante production. Elles demandent de nouvelles approches (Lenartz, 2020 ; Makhlouf Shabou, 2023a).

[3] « 3.1.4. information (3.1.3) that has value to the relevant stakeholder », ISO 24143
« Information, data, documents, or records created or received by an organization or person in any format », InterPARES Trust AI
[4] Une section de la revue de la littérature a été consacrée à la compréhension du vrac numérique et à la recherche de termes équivalents en anglais (partie 1.4.1)
[5] Le versionnage des documents inclus les versions multiples d'un même document : brouillons, copies, versions finales ou encore originales.

1.3.5. Données non structurées — Unstructured data (EN)

Données stockées sans organisation particulière, tant par leurs sources, formats ou classement « ce qui rend leur utilisation difficile dans un système d'information » (Grand dictionnaire terminologique, 2023). La plupart des options de recherche classiques ne leur sont pas facilement applicables (Lenartz, 2020). Ces données comprennent le plein texte et l'audiovisuel.

1.3.6. Évaluation — Appraisal (EN)

Étape permettant de juger de la valeur des documents afin de déterminer leurs périodes de conservation (Couture, 2011). Les services producteurs sont responsables de la sélection des documents produits selon leur valeur qu'ils sont les seuls à pouvoir discerner avec justesse. Il est nécessaire d'intégrer les principes de temps, méthode, participation, documentation et guides pratiques (The National Archives, 2022). Ces derniers permettent la sensibilisation, la responsabilisation et l'intégration de l'évaluation dans les processus de travail, directement au sein des équipes productrices.

1.3.7. Automatisation — Automation (EN)

Processus par lequel des tâches ou des opérations sont réalisées grâce à l'utilisation d'algorithmes ou de machines. Il vise à réduire ou éliminer l'intervention humaine afin d'accroître l'efficacité et la précision. Dans le domaine du traitement des données, l'automatisation contribue à réduire les délais et à optimiser la prise de décision en permettant aux professionnel·le·s de se concentrer sur des tâches plus complexes nécessitant des compétences spécifiques (Makhlouf Shabou et al., 2020, p. 183 ; Manaher, 2021 ; Merriam Webster, 2023).

1.3.8. MPLP — More Product Less Process

Processus défini par Greene et Meissner qui vise à aider les spécialistes documentaires dans leurs tâches. Il est basé sur quatre principes : rendre les collections accessibles dans les meilleurs délais, assurer la mise à disposition adéquate des documents en fonction des besoins des usagères et usagers, appliquer des mesures minimales pour la préservation physique des documents, et finalement décrire suffisamment les documents à des fins d'utilisation et de consultation. Malgré les critiques concernant la simplification du processus de traitement archivistique (Phillips, 2015) et l'ambiguïté sur la manière de prioriser le traitement des documents (Mersiovsky, 2014), le MPLP a ouvert la discussion sur l'adaptation nécessaire du traitement pour les larges volumes de documents en gardant à l'esprit l'accessibilité aux collections (Mersiovsky, 2014).

1.3.9. Archiving by design

Concept mettant l'accent sur l'intégration proactive de la gestion des archives, dès la conception initiale de l'information. Il s'agit de prendre en compte les exigences de préservation des documents, de classification et la gestion des métadonnées dès le départ, afin de garantir que les informations cruciales soient capturées, organisées et préservées de manière efficace. Six principes d'accessibilité durable sont au cœur d'archiving by design : trouvable, disponible, lisible, interprétable, fiable et à l'épreuve du temps[6] (EAG — European

[6] « Findable, available, readable, interpretable, reliable, future proof ».

Archives Group 2023). Cette approche durable permet de réduire les risques liés à la perte de données importantes et de faciliter l'accès futur aux informations (Hoolt, 2023). Cependant, s'agissant d'une approche conceptuelle, elle n'offre ni de marche à suivre ni de solution clé en main qui soit applicable à toutes les situations.

2. Méthodologie

Dans le chapitre ci-dessous, nous allons détailler les différentes étapes de notre méthodologie afin de permettre la reproductibilité de notre démarche et la compréhension des biais potentiels.

2.1. Veille et revue de la littérature

Nous avons mis en place une veille informationnelle à l'aide de mots-clés[7] sur Google alert, complétée par deux outils de cartographie de publications scientifiques (Inciteful et ResearchRabbit).

Notre approche étant expérimentale et descriptive, nous avons procédé à un état des lieux élargi sur la recherche concernant les vracs numériques et les données de masse non structurées. L'état de la recherche sur la prise en main de vracs numériques audiovisuels à des fins de valorisation est encore immature. Cependant, le nombre croissant de publications souligne un intérêt grandissant et des besoins urgents de la communauté scientifique. Durant les quelques mois de notre recherche, notre veille est venue confirmer nos résultats. Les résultats de notre revue de la littérature sont répartis en cinq sous-points thématiques.

2.1.1. Vrac numérique : caractéristiques et prise en main

Peu d'articles scientifiques et de travaux académiques mentionnent explicitement les vracs numériques, bien que de nombreux travaux introduisent des concepts de gestion de données à l'ère du numérique et de la production de masse dans une optique de maîtriser ces données non structurées[8].

Dans la littérature anglophone nous retrouvons l'utilisation de data swamp qui illustre un ensemble hétérogène, méconnu, peu ou non maîtrisé, qui recèle un potentiel peu ou pas exploité et qui est en l'état impraticable (Koch, 2018). Les caractéristiques les plus communes des data swamps sont (Atlan, 2023 ; « Data Swamp », 2023) :

- Le manque de métadonnées
- La mauvaise qualité des données
- Le manque de gouvernance des données
- La présence de problèmes en lien avec la sécurité et les normes
- La mauvaise utilisation des ressources

[7] Mots-clés utilisés : archival appraisal ; appraisal + 'unstructured data'; archiv* + automation + appraisal, automation appraisal; automatisation + « archiv* » ; automatisation « donnée* non structurée* » ; automatisation + « vrac numérique »
[8] Nous en avons vu un exemple à travers la notion d'archiving by design (EAG — European Archives Group 2023; Hoolt 2023).

- La volumétrie

Ces caractéristiques rejoignent celles des vracs numériques définis par l'Association des archivistes français présentées dans la partie 1.3.3. Dans les deux termes, la notion de volumétrie est capitale. Le potentiel informationnel est inexploitable par les institutions en raison d'une gestion documentaire inadaptée, lacunaire ou absente. Malgré l'importance grandissante de la prise en main de ces actifs inexploités, les stratégies existantes pour l'analyse et la gestion des vracs manquent d'adaptabilité aux réalités pratiques des organismes et des utilisateur·rice·s (Bischoff, 2022).

Belovari nous offre un point de vue pratique sur un vrac audiovisuel qu'elle définit comme étant un volume de données non structuré (Belovari, 2019). Bien que la taille du fonds évalué par l'autrice — 677 Go — ne soit pas comparable à celui que nous allons traiter — 2,34 To —, sa recherche nous a confirmé l'application d'un traitement automatisé pour certaines étapes de l'évaluation. Belovari comme Texier recommande de procéder au traitement des données au niveau macro (broad appraisal) et de mener une enquête sur le contexte de création afin de comprendre la structure du vrac et l'état de santé des données stockées (Belovari, 2019 ; Texier, 2021a). L'automatisation assiste la·le spécialiste dans l'identification des premiers problèmes qui créent du bruit autour des données : doublons, formats obsolètes, fichiers corrompus ou vides, etc. L'étape suivante consiste à opérer une évaluation au niveau des documents (in-depth appraisal) durant laquelle surgissent les questions de conservation et de valorisation patrimoniale. Ainsi, la priorité du traitement va aux dossiers, puis à l'évaluation de l'item. Cependant, cette étape n'est à faire que lorsqu'elle est possible ou nécessaire parce que fortement chronophage (Belovari, 2019 ; Bussard, 2022 ; The National Archives, 2022).

Le travail de Bussard réalisé la Cinémathèque suisse nous a apporté des pistes d'automatisation pertinentes sur la recherche de redondances strictes, la visualisation de l'arborescence et l'identification de la volumétrie par type (Bussard, 2022). Le travail de Bennani, Hategekimana et Rey Rodriguez sur les pratiques dans différentes institutions suisses démontre que les volumes de données non textuelles atteints sont difficiles à gérer pour les institutions (Bennani et al., 2022). La « massification [des données] impose aux institutions patrimoniales de revoir leurs critères d'acquisition, leurs méthodes de gestion et d'examiner les possibilités de concertation et de mutualisation des engagements » (Coutaz, 2018, p.32). Avec l'automatisation, il s'agit de trouver des méthodes efficaces de traitement des données qui ne peuvent plus être traitées avec des méthodes traditionnelles. Les travaux de Bischoff et Makhlouf Shabou soulignent que les stratégies existantes pour l'analyse et la gestion des vracs numériques manquent d'adaptabilité aux réalités pratiques des organismes et des utilisateur·trice·s (Bischoff, 2022 ; Makhlouf Shabou, 2023a).

2.1.2. Archivage numérique : normes et bonnes pratiques

Plusieurs normes ISO (International Standardisation Organisation) clarifient les documents à conserver et rappellent les avantages d'une gestion documentaire adaptée. Celle sur la gouvernance informationnelle émet quinze principes pour atteindre une gouvernance optimisée (Organisation Internationale de Normalisation, 2022b). Elle souligne l'importance de valoriser toute l'information produite en tant qu'actif informationnel (partie 1.3.2). Elle soutient également la nécessité d'utiliser une approche de gestion basée sur les risques, comme l'ISO 15489-1 (Organisation Internationale de Normalisation, 2016). Les documents à conserver doivent être authentiques, intègres et fiables (Organisation Internationale de Normalisation,

2018, p. vii). Les normes stipulent toutes que les documents doivent être maîtrisés au plus tôt, ce qui signifie :

- Identifier clairement les producteur·rice·s et les propriétaires des données,
- Capter les informations liées au contexte de création et aux pratiques,
- S'assurer de la qualité et de l'intégrité des données

Dans les guides pratiques de plusieurs institutions suisses et étrangères, ainsi que dans des articles de spécialistes, nous relevons que les documents numériques sont à traiter avec une approche ascendante afin de ne pas être débordés par la masse de données ou leurs journaux d'arriérés (backlogs) (Bunn, 2023 ; M. A. Greene, 2010). L'archivage numérique s'inspire de l'approche d'évaluation MPLP (partie 1.3.8), qui identifie la mise à disposition la plus rapide possible des documents au public comme le principal challenge des archivistes (M. A. Greene, 2010 ; M. Greene & Meissner, 2005). Plusieurs guides reviennent sur l'importance de la responsabilisation des producteur·rice·s de documents. Seuls elles et ils sont les plus à même de juger de la valeur et de l'importance des éléments générés, ainsi que les connaissances pour enrichir les informations sur le contexte de création (Conférence des recteurs et des principaux des universités du Québec, 2009 ; The National Archives, 2013). Le rôle des archivistes est de renseigner les producteur·rice·s (Association des archivistes français, 2018 ; Coutaz, 2016a ; The National Archives, 2013, 2022).

2.1.3. Évaluation automatisée des données de masse : perspectives et challenges

La recherche de l'évaluation automatisée de masse de données audiovisuelles est en phase exploratoire. Des outils intéressants sont développés par des archives nationales, des gouvernements ou encore des compagnies privées. Ils facilitent les tâches des professionnels dans l'analyse des contenus volumineux ou non structurés (Bavaud et al., 2022 ; Oguey & Schneiter, 2018) afin de permettre une meilleure efficacité de traitement, la maîtrise de la volumétrie, la localisation effective des actifs informationnels, ce qui vient directement contribuer à l'amélioration de la prise de décision et de la recherche scientifique (Makhlouf Shabou, 2023b). Cependant, l'utilisation de ces outils est fortement tributaire de la qualité des données d'entrée — leur structuration et propreté — et représente une perte de flexibilité pour les professionnels (Aas, 2018).

L'automatisation ne remplace pas l'intervention humaine. La compréhension des besoins et l'identification des outils sont deux étapes capitales dans la mise en place d'une automatisation efficiente (Association des archivistes français, 2018). Cependant, les professionnel·le·s ne bénéficient pas, à ce jour, de bases de données recensant les différentes options d'automatisation (Bavaud et al., 2022). Nous notons qu'il existe un tableau panoptique en annexe du travail de mémoire de Bavaud, Bischoff et Bussard publié en 2022.

Le recours à l'automatisation doit intégrer des questions d'éthique, de qualité, de représentabilité des données d'entrée, ainsi que la vérification et le contrôle des résultats obtenus. L'article de Aas sur la NAE (Archives nationales d'Estonie) nous fournit un exemple intéressant sur ces questions (Aas, 2018). Dans son article, Aas illustre l'importance d'accompagner l'automatisation d'une vérification humaine des décisions prises par les algorithmes utilisés afin d'éviter des descriptions erronées ou des erreurs dans les droits d'accès. L'intervention humaine est nécessaire tout au long de l'automatisation afin de garantir

l'intégrité des données d'entrée, la mise en place des critères de sélection justes et la conformité des données en bout de chaîne (Makhlouf Shabou et al., 2020).

2.1.4. Valeur patrimoniale

Dans sa recherche sur le fonds audiovisuel d'une institution scolaire, Belovari propose quelques pistes de réflexion pour l'évaluation in-depth, c'est-à-dire l'évaluation au niveau de l'item, utile à la valorisation patrimoniale. L'autrice isole trois types de critères de sélection : personnel, relatif au contenu et relatif à l'esthétique (Belovari, 2019, p. 65). Le premier critère est en lien étroit avec la sensibilité et l'expertise acquise par l'archiviste. L'expérience du·de la professionnel·le vient alimenter et affiner les capacités d'évaluation. Les critères de contenu renvoient aux documents faisant preuve de « nouveauté » ou d'« unicité » (Belovari, 2019, p. 66). Les critères d'esthétique encouragent l'archiviste à privilégier la qualité du support, mais également de l'angle de vue[9]. Finalement, tous ces critères sont croisés avec les critères légaux — comme la protection des données personnelles —, les conditions de conservation et de stockage, et finalement les durées de conservation (Belovari, 2019, p. 64).

Une problématique récurrente lorsque l'on aborde les archives patrimoniales audiovisuelles est celle du support et de son évolution dans le temps (Bussard, 2022 ; Memoriav, 2023). L'enjeu est de préserver, à travers la numérisation notamment, les supports des photographies ou des vidéos analogiques qui sont non seulement un médium, mais également un document en soi (Lefort, 2018). Le travail de Lefort nous éclaire sur certains critères à considérer lors de la sélection de documents patrimoniaux, qui rejoignent la catégorie des critères esthétiques de Belovari : qualité du support, rareté du sujet ou de la technique, qualité de la mise au point, niveaux d'exposition, etc. Lefort recommande l'implémentation d'un échantillonnage pour les fonds volumineux (Lefort, 2018, p. 27-28).

La recherche de Scarpulla, portant sur les archives des arts vivants, souligne les problématiques liées à la valeur patrimoniale, illustrée par le lien étroit entre la captation de performances — le témoignage d'une pièce passée — et la reproduction des œuvres captées — la recréation. Ces captations forment des « traces » qui sont consultées, imitées, reproduites, recréées (Scarpulla, 2016) et sont ensuite « fragmentées, réduites en unités d'information — les données — pour devenir manipulables » (Bardiot, 2022, p. 312). Ces archives doivent prendre en compte plusieurs critères patrimoniaux :

- Le témoignage de l'activité,
- Le potentiel de reproductibilité d'une pièce[10],
- La valorisation des activités des compagnies et des écoles,
- La promotion et la diffusion des évènements sur le web,
- L'exposition ou la présentation au public et

[9] Dans le cadre de son cas pratique, Belovari va privilégier des prises de vue des étudiant·e·s de profil afin de montrer leurs appareils auditifs et documenter l'évolution de cette technologie.
[10] Dans le cadre des arts vivants, une pièce est toujours activable. C'est-à-dire qu'à partir du moment de sa création et de sa première présentation, elle peut toujours continuer à être jouée et présentée, que ce soit par l'artiste ou par d'autres interprètes. Chaque utilisation à des fins reproduction, formation, promotion ou transmission participe au « prolongement créatif » de l'œuvre (Scarpulla, 2016, p. 23), la maintenant dans un statut actif.

- L'utilisation à des fins de formation.

Nous constatons que bien souvent, les archives patrimoniales conservées et mises en avant sont celles provenant des enseignements ou des performances, et « non pas celles des équipes techniques ou administratives » (Scarpulla, 2016, p.31), ce qui n'est finalement que peu représentatif de la complexité des activités d'une institution. Belovari arrive à une conclusion similaire : elle souligne qu'en l'absence de contrôle de la masse documentaire de la part des productrices et producteurs, les archivistes sont amenés à mettre en place des priorités au sein des collections et des créateur·rice·s de documents (Belovari, 2019). Cette priorisation risque de dévaluer la représentativité des fonds.

2.1.5. État de l'art des archives administratives

Les Archives fédérales suisses (AFS) nous indiquent que les documents issus des institutions publiques et destinés à l'archivage à long terme sont versés aux archives publiques, cantonales ou fédérales. De manière générale, le cycle de vie des documents courants est délimité par un plan de classement et un calendrier de conservation qui permettent de déterminer le sort final des documents. Selon l'AFS, celui-ci doit être connu dès la création, même s'il est parfois question d'évaluation rétrospective selon des critères que nous n'avons pas pu consulter (AFS, 2022).

La Loi sur l'Archivage définit les archives publiques comme étant le résultat organique de la production d'une institution reliée à leur usage courant (LArch, art.3) (République et canton de Genève, 2000b). Une distinction est faite avec les archives historiques[11] conservées en priorité dans les archives publiques (LArch, art.5). Y sont archivés les documents ayant une valeur juridique, politique, économique, historique, sociale ou culturelle (LArch, art.2), faisant partie du domaine public (LArch, art. 2).

Les services administratifs de droit public à Genève ont l'obligation de verser leurs archives définitives aux Archives d'État (République et canton de Genève, 2000 a, art.7). La libre consultation des archives publiques est garantie par la loi (République et canton de Genève, 2000 a, art.11). En résumé, les institutions de droit public ont pour obligation de maîtriser leurs documents, de les verser aux archives et de les mettre à disposition dès que possible selon les délais définis par la loi (République et canton de Genève, 2000 a, art.12).

2.2. Analyse des besoins

Ce projet de recherche a été effectué sous mandat pour le service communication d'une institution de droit public. Ce service est actif depuis quinze ans et a subi de nombreux changements sur les trois dernières années : arrivée de collaborateur·rice·s, départs d'anciens membres, nouvelle direction, déménagement et nouvelle ligne graphique. Il est composé de quatre membres, ponctuellement assistés par des prestataires externes pour certains services (captation, campagnes d'affichage, etc.). L'équipe est responsable du site internet de l'institution, des réseaux de communication interne, des newsletters, réseaux sociaux, campagnes d'affichages, bande-annonce et toute autre promotion ou production documentaire concernant plus de 250 évènements annuels.

[11] Les archives historiques sont « l'ensemble des documents qui ne sont plus utiles pour l'expédition courant des affaires et qui sont conservés en raison de leur valeur archivistique définie par les principes et dispositions de la présente loi » (LArch, art.3).

2.2.1. Rencontre avec les mandants

Nous avons effectué deux entrevues avec notre mandant : un kick off afin de s'accorder sur les objectifs, expliciter la méthodologie et établir les grandes lignes du calendrier pour convenir des prochaines étapes, et un entretien pour répondre à des précisions, visiter les locaux et lancer les analyses automatisées.

En prévision du deuxième entretien, nous avons transmis en amont la liste de nos questions ainsi que les instructions des logiciels d'analyse Archifiltre-Docs[12] et DROID. L'installation sur une machine du service nous a permis de garantir que les données restent au sein du service. Nous avons récupéré la « photographie » de l'arborescence et de la santé des données à un instant T.

Les informations récoltées durant les entretiens sont structurées en trois catégories : l'état des lieux du parc informatique (partie 2.2.2), les principes directeurs (partie 2.2.3) et les pratiques courantes du service (partie 2.2.4).

2.2.2. État des lieux du parc informatique

L'institution comprend plusieurs bâtiments à Genève, avec une structure informatique conséquente. Elle souhaite rationaliser ses ressources et améliorer son empreinte écologique. Tout le parc informatique est centralisé dans un même datacenter avec un service IT dédié. Le service communication a accès à un serveur consacré aux documents d'activité courante. Un autre serveur (cold storage) contient des projets terminés, versés dans l'optique d'une évaluation à des fins d'archivage, de destruction et de désengorgement du serveur courant. Cependant, suite aux divers changements opérés au sein du service communication, ce cold storage est resté à l'état de vrac et a été oublié. Ce phénomène est également présent dans d'autres services.

Sur site, un service informatique est spécifiquement responsable des installations et du lien avec le datacenter. L'équipe de communication fonctionne sur un environnement mixte Windows et macOS. Chaque employé possède un poste de travail dédié. Les employé·e·s travaillent de manière collaborative soit sur le serveur des activités courantes, soit sur des services cloud partagés par l'équipe.

Notre mandat porte exclusivement sur le cold storage de l'équipe de communication qui devrait ne contenir que des projets et documents clos, inutiles aux activités courantes.

2.2.3. Principes directeurs de l'institution publique

L'institution a produit différentes directives dont les plus importantes pour nous sont : la gouvernance documentaire, le plan de classement et la procédure de traitement des dossiers inactifs. Toutes ces directives datent des 5 dernières années.

Nous retenons de la directive de gouvernance que :

- Les rôles et responsabilités doivent être définis,
- Chaque membre du personnel et prestataire de service est responsable de l'application de la directive,

[12] Pour plus de commodité pour la suite de la lecture de cet article, nous utiliserons dorénavant le terme « Archifiltre » et non « Archifiltre-Docs ».

- Chaque projet doit donner lieu à un dossier qui rassemble l'ensemble cohérent des documents relatifs et
- Tous les documents doivent répondre aux principes d'intégrité[13] et d'authenticité[14].

Les responsables désignés sont tenus, au terme de la durée de conservation, de proposer les dossiers pour l'archivage historique ou la destruction, conformément à ce qui est indiqué dans le plan de classement.

Le plan de classement catégorise et réglemente les documents. Ceux-ci intègrent la production documentaire du service communication, lui-même situé au sein des services communs et administratifs. Le service communication est responsable du classement de huit des catégories[15]. Seule la moitié d'entre elles sont à archiver. Il s'agit des documents relatifs à la stratégie, la création et la diffusion, la gestion des images, les manifestations et le marketing publicité.

La procédure des documents inactifs détaille la marche à suivre pour ce type de document. Notons que l'élimination de dossiers doit être validée par un bordereau de destruction. Les Archives de l'État de Genève (AEG) n'ont pas reçu de versement de documents par manque de place dans leurs locaux. À l'heure actuelle, tous les documents sont conservés au sein de l'institution.

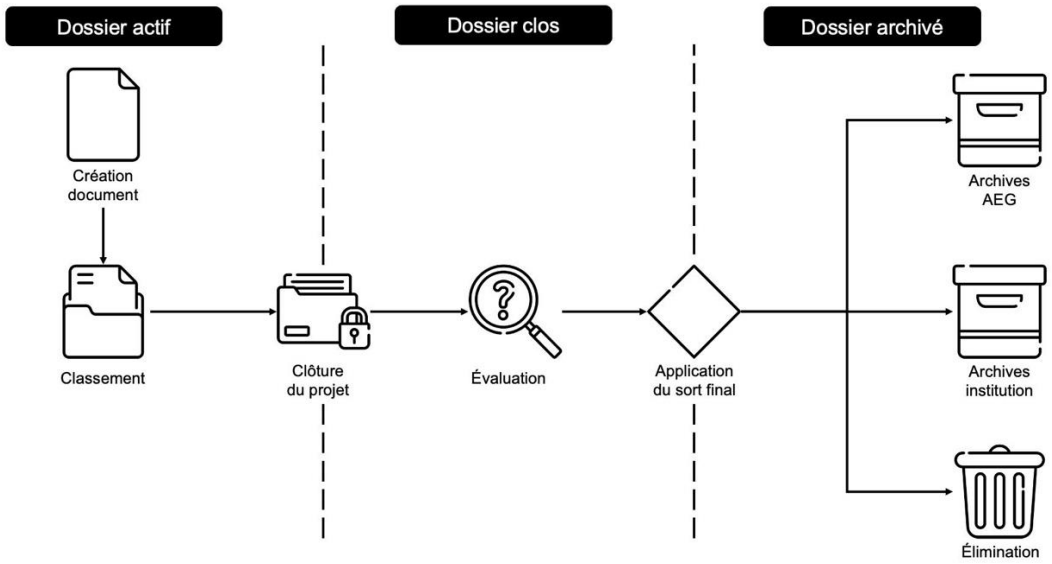

Figure 1 : Processus de traitement des documents

(Autrices)

À la lumière des principes directeurs, nous identifions trois problèmes :

[13] Le document se doit d'être complet et non altéré.
[14] Le document est bien ce qu'il prétend être, en d'autres termes qu'il a été produit ou reçu par la personne qui prétend l'avoir produit ou reçu, et qu'il a été reçu au moment où il prétend l'avoir été.
[15] Les huit catégories sont la responsabilité du service communication sont : stratégie, gestion des contacts, identité visuelle, promotion, création et diffusion, gestion des images, gestion des partenaires, manifestations et marketing publicité

- Certaines catégories du plan de classement méritent d'être clarifiées, par exemple la catégorie « Gestion d'images » qui est explicitée par « Conserver les photos importantes pour l'histoire de l'institution » sans fournir plus de détails sur les critères à prendre en compte,
- Certains types de documents produits par le service communication ne sont pas inclus dans les descriptifs du plan de classement, notamment les vidéos promotionnelles (trailers) qui devraient être incluses dans « Production de contenus informationnels »,
- La destination des archives historiques n'est pas claire.

2.2.4. Pratiques du service communication

Par année, le service communication suit 250 évènements. La grande majorité est récurrente et se reproduit tous les ans ou les deux ans. Pour certains, l'équipe fait appel à des prestataires externes pour la promotion à plus grande échelle ou pour la captation des évènements. Ces derniers transmettent les rushs et autres documents de travail brut qui sont rangés dans le dossier du projet.

Chaque année, le service reçoit en moyenne cinq demandes externes d'accès à des documents auxquelles il doit répondre. Cependant, il peine à effectuer cette tâche à cause de la difficulté de navigation dans le cold storage et par manque de temps.

L'équipe classe de manière générale les dossiers par année puis par projet. Lorsqu'un projet est terminé, l'équipe passe directement au suivant, sans intervenir ou trier le projet qui a été clôturé. Les documents sont produits de manière collaborative, directement sur le serveur des activités courantes ou sur le cloud. Le service communication utilise couramment plusieurs logiciels propriétaires, tout particulièrement la suite Adobe (Photoshop, Illustrator, InDesign, Premiere Pro) et Microsoft (Word, Excel, PowerPoint). L'intégralité des documents produits ou reçus sont conservés en l'état « au cas où ». Aucune conversion de format n'est effectuée sur les documents après la clôture du projet.

Les plus anciens membres de l'équipe transfèrent les dossiers inactifs dans le cold storage afin de libérer le serveur courant. L'autre objectif de cette étape est à terme d'évaluer les dossiers de projets passés pour l'archivage ou la suppression. Malheureusement, les données ne semblent pas avoir été évaluées et donnent lieu au vrac numérique que nous allons analyser. Le vrac ne semble pas structuré selon le plan de classement.

À terme, le serveur hébergeant le cold storage a vocation à être désactivé.

2.3. Risques encourus

Le plus grand risque, très réel, que le service court à l'heure actuelle est la perte des données en raison de l'introuvabilité des informations, l'inexploitabilité du vrac ou encore par la désactivation des serveurs obsolètes. En effet, le cold storage oublié du service a failli être supprimé lors du dernier audit informatique.

Nous détectons également d'autres risques et menaces :

- Difficulté de navigation,
- Corruption des données,
- Formats obsolètes et plus pris en charge,
- Non-conformité des délais de conservation selon la législation,

- Non-respect de la protection des données personnelles.

Ces derniers pourraient impacter la réputation de l'institution en cas de cyberattaque ou de fuite.

Nous notons également qu'il n'existe aucun inventaire des projets hébergés dans le cold storage en cas d'accident. Tous ces éléments viennent confirmer les risques propres au vrac numérique (partie 1.4.1). La perte de ces données aurait un impact non seulement sur le fonctionnement du service, mais aussi pour le patrimoine historique de l'institution.

2.4. Choix et application des logiciels

Comme cité dans la revue de la littérature (partie 1.4), il n'existe aucun programme qui permette de traiter toutes les fonctions archivistiques (Bavaud et al., 2022). Nous avons opté pour une combinaison de programmes. Chacun effectue un scan des documents et fournit une image à un instant T du fond analysé. Dans l'optique d'une généralisation de notre processus et afin de favoriser la reproductibilité de notre recherche, nous nous sommes concentrées sur des programmes gratuits et open source.

Les logiciels listés fonctionnent tous avec la base de données de formats de fichiers réalisée par la National Archives du Royaume-Uni : PRONOM. Cette base de données recense plus de 1 400 formats et est régulièrement mise à jour.

2.4.1. DROID

DROID (Digital Record Object Identification) est développé par les National Archives du Royaume-Uni. Il sert à l'identification des formats de fichiers en fournissant plusieurs paramètres comme le type, la date de modification, les extensions, les extensions erronées, les versions ou encore la taille (The National Archives, 2023).

2.4.2. Archifiltre

Archifiltre a été développé par une start-up d'État faisant partie de la Fabrique Numérique des Ministères Sociaux (Archifiltre, 2023). Lancé en réponse à la nécessité de gérer efficacement des volumes de données grandissantes. La grande force du logiciel réside dans la visualisation de l'arborescence et des filtres qui peuvent lui être appliqués. Combinés, ces éléments fournissent rapidement une vue d'ensemble de la structure et de la volumétrie.

2.5. Récapitulatif de l'état des lieux du mandant

Le vrac numérique est le résultat des divers changements opérés dans le service communication (partie 2.2), et d'une transmission lacunaire du savoir tacite des différents membres de l'équipe. Le cold storage semble être constitué de projets terminés. Il est hébergé sur un serveur temporaire en attente d'évaluation pour l'application des sorts finaux des documents. L'équipe utilise de nombreux logiciels propriétaires sous licence (Adobe Creative Suite et Microsoft Suite). La production documentaire du service est composée des multiples formats qu'ils produisent à l'interne ainsi que des fichiers bruts reçus des prestataires externes. Tous les documents reçus ou produits sont conservés en l'état. L'équipe n'a pas intégré dans son workflow une période dédiée à l'évaluation des fichiers, et nous fait part du manque de temps alloué à cette tâche. À l'heure actuelle, les données sont à risque de perte et nécessitent un traitement.

3. Description des données

La description des données est effectuée sur l'appui des documents extraits des logiciels DROID — un fichier .csv sans empreintes, un rapport (Comprehensive breakdown) — et Archifiltre — un fichier .csv avec empreintes, un rapport (Rapport d'audit) et un fichier. json. Nous nous sommes assurés de la conformité des données en vérifiant que les fichiers étaient lisibles, et en effectuant un petit échantillonnage aléatoire. Pour cela nous avons retenu vingt fichiers et dix dossiers et comparé leurs chemins identifiés par chaque programme. Cette étape nous a permis de confirmer que le récolement des données des deux programmes avait bien fonctionné.

Pour des éléments comparables, les logiciels nous ont fourni des chiffres différents, bien que suffisamment proches. Cela est partiellement expliqué par la granularité de détection de formats et problèmes liés de chaque programme. Cependant, les résultats fournis sont complémentaires et nous offrent une meilleure vision du vrac.

3.1. État du disque

Le Rapport d'audit d'Archifiltre identifie 33 418 fichiers dans le vrac réparti dans 832 dossiers. Le volume total de l'arborescence est de 2,6 To. Le Comprehensive breakdown de DROID présente des chiffres légèrement différents. Il comptabilise un total de 37 033 éléments, dont 36 095 sont des fichiers. Pour identifier le nombre de dossiers, nous avons soustrait le nombre de fichiers au nombre d'éléments totaux (37 033 - 36 095). Nous trouvons 938 dossiers. Le Tableau 2 compare les résultats obtenus avec les deux logiciels et nos propres chiffres.

En ce qui concerne les dates extrêmes, les programmes utilisent des méthodes similaires de tri. Les résultats obtenus sont basés sur les dates des documents (ou éléments) et n'incluent pas les répertoires (ou dossiers) (Figure 2 et Figure 3).

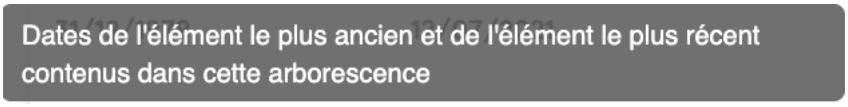

Figure 2 : Méthode utilisée par le logiciel Archifiltre pour les dates extrêmes

(Archifiltre)

Report field	Grouping fields	
FILE_SIZE	Year(LAST_MODIFIED_DATE)	
Filter fields:		
Field	**Operator**	**Values**
RESOURCE_TYPE	NONE_OF	"Folder"

Figure 3 : Méthode utilisée par le logiciel DROID pour les dates extrêmes

(DROID)

Archifiltre présente les dates au format JJ/MM/AAAA. Le fichier le plus ancien est daté au 31 décembre 1979 et le plus récent au 12 juillet 2021.

DROID situe les dates extrêmes du vrac en 1980 et 2021, avec 11 fichiers dont la date n'a pu être déterminée. Ce logiciel nous fournit trois options de granularité plus fine :

- Par année au format AAAA,

- Par année et par mois, au format AAAA suivi du chiffre du mois,
- Et par mois de modification, avec le chiffre du mois.

Grâce à ces options, nous pouvons voir quels sont les mois (Figure 4) et les années (Figure 5) qui ont comporté le plus grand nombre de dernières modifications de documents. Ainsi les mois de mai et d'octobre sont les plus actifs avec 42,87 % des documents présents sur le vrac — soit 15 476 éléments —, suivis par les mois de septembre et d'août qui comptabilisent 19,84 % des fichiers — soit 7 163 éléments.

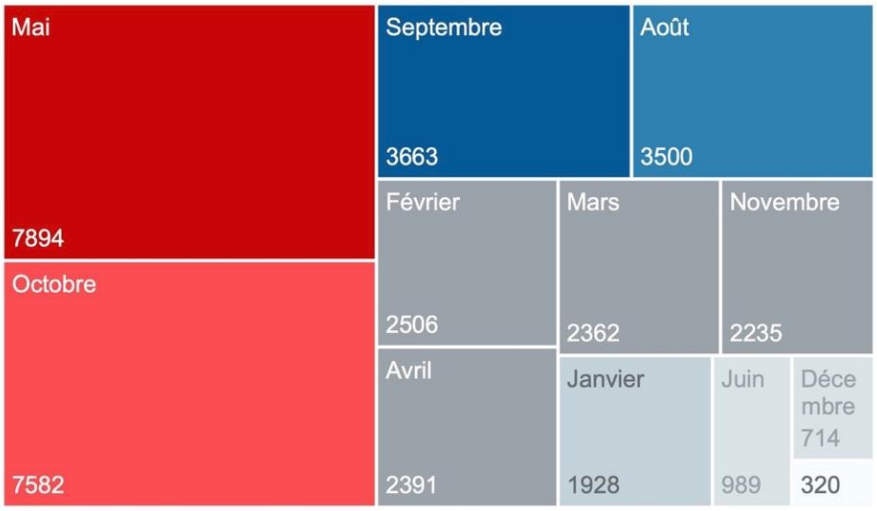

Figure 4 : Nombre de documents modifiés pour la dernière fois par mois, toutes années confondues

(Autrices)

Du côté des années les plus actives (Figure 5), 73,18 % des données ont été modifiées pour la dernière fois entre 2009 et 2016. Deux pics sont enregistrés en 2013 (11 460) et en 2015 (6 150). Le vrac contient 23,25 % de documents ultérieurs à la création de l'institution, soit 8 394 fichiers. Finalement, moins de 4 % des données ont été modifiées pour la dernière fois après 2017, ce qui représente 1 273 documents.

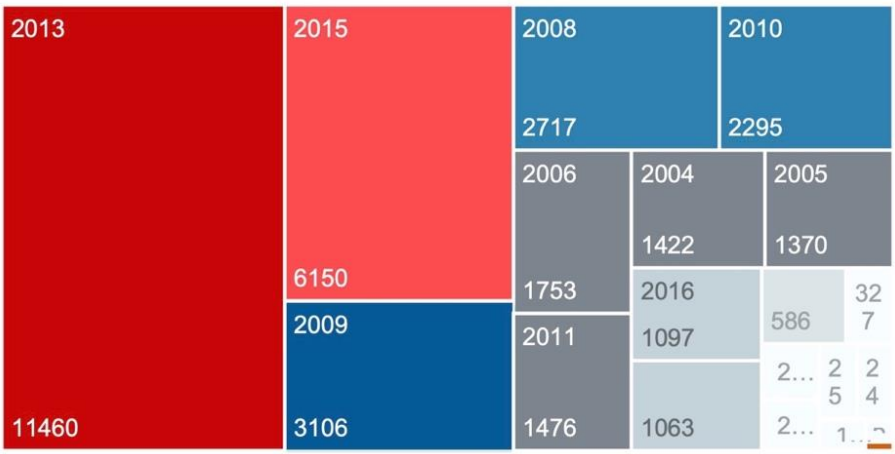

Figure 5 : Nombre de fichiers modifiés pour la dernière fois par année, tous mois confondus

(Autrices)

3.2. Arborescence

L'arborescence du vrac comporte 13 niveaux à son point le plus profond. La moyenne de profondeur se situe entre 4 et 5 niveaux. Dans la Figure 6, les dossiers sont représentés en jaune et en bout de visuel nous retrouvons des fichiers : les vidéos en violet et les fichiers audio en rose. Cette visualisation est pondérée par volume, ce qui signifie que la taille des éléments est proportionnelle au poids des documents contenus.

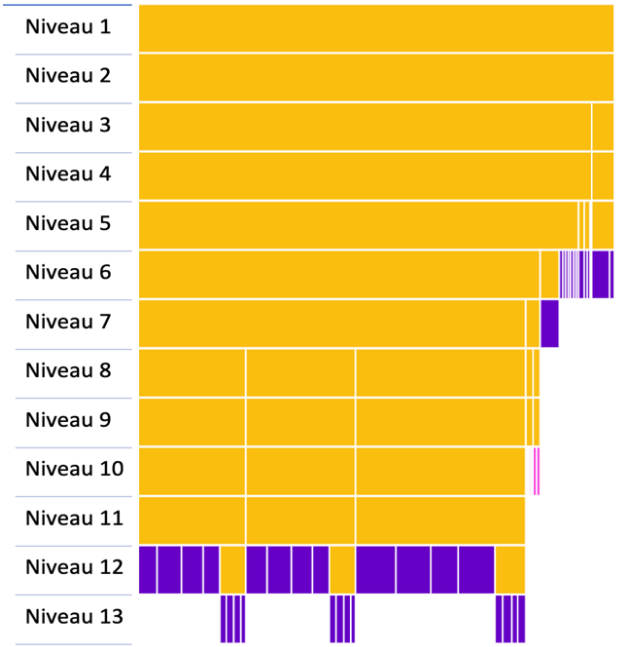

Figure 6 : Extrait de l'arborescence avec visualisation des niveaux de profondeur

(Archifiltre)

3.3. Diagnostic des données

Nous allons nous focaliser sur les fichiers qui créent du bruit autour des données. Ceux-ci sont des documents éliminables à tout moment par l'équipe, sans nécessiter l'aval d'un·e archiviste pour la suppression (Belovari, 2019 ; Texier, 2021a). Il s'agit des doublons, fichiers temporaires et fichiers corrompus.

Les doublons sont identifiés par des hashs[16] qui sont une empreinte unique des fichiers sous forme de chaîne de caractères (Office québécois de la langue française, 2023). Les doublons comportent entre 2 et 8 exemplaires au sein du vrac.

Le volume total des doublons est de 64,4 Go, comprenant 8 402 éléments. Archifiltre segmente ce volume en trois catégories : Image, Autre et Vidéo. Image contient le plus d'éléments avec 7 884 fichiers qui occupent 25,5 Go des redondances. Autre contient 495 éléments pour un poids total de 12,3 Go. Finalement, Vidéo ne contient que 13 éléments pour un volume total de 26,2 Go. Dans la figure ci-après exprimée en pourcentage, nous illustrons la relation entre le nombre de fichiers redondants identifiés par type et leur volume. Ainsi, les vidéos, qui

[16] Archifiltre utilise l'algorithme de hachage MD5 et DROID en utilise trois : HA 1, SHA 2 (256) or MD5 (The National Archives, 2023, p. 12)

représentent moins de 1 % des éléments redondants, occupent à elles seules 41 % du volume total des doublons, soit virtuellement le même poids que les 7'8884 images redondantes.

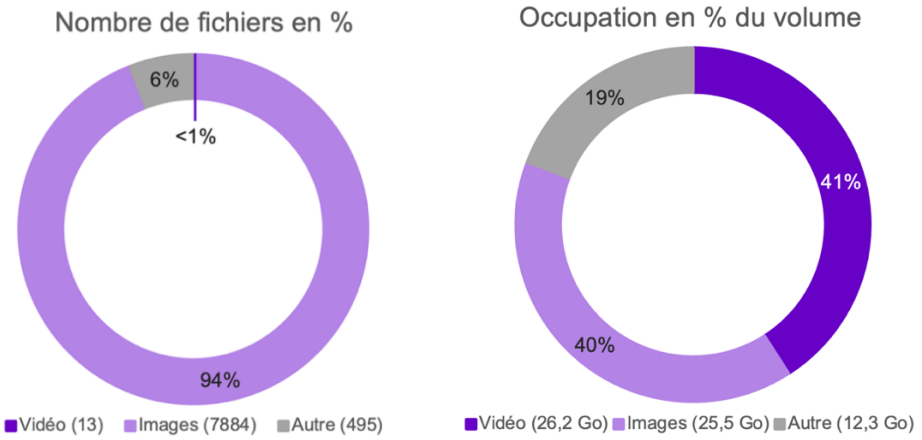

Figure 7 : Comparatif de la répartition totale des doublons par type, en nombre de fichiers et en occupation

(Autrices)

Les fichiers temporaires sont des fichiers créés par les systèmes d'exploitation ou les logiciels utilisés par le service communication. Ces éléments utiles dans le cadre de données actives et des projets en cours deviennent inexploitables une fois ceux-ci clos. Ils deviennent des fichiers fantômes qui occupent inutilement de la place sur le serveur. Les logiciels Archifiltre et DROID ne permettent pas directement l'identification des fichiers temporaires. Cependant, grâce à l'analyse des extensions avec PRONOM et des bases de données informatiques (File Extension, 2023 ; The File Format Database, 2023; The Source for File Extension Information, 2023), nous avons identifié 960 éléments (Tableau 1).

Extension	PUID attribué par DROID	Nombre de fichiers identifiés
.ds_store	fmt/503	172
.db	fmt/111, fmt/682	37
.pkf		13
.bdm	fmt/1075, fmt/1076	6
.bridgecachet		96
.bridgecache		82
.out	fmt/101	25
Sans extension	fmt/503[16]	529

Tableau 1 : Nombre d'éléments temporaires présents sur le vrac, par extension

(Autrices)

Finalement, les fichiers corrompus sont des documents dont l'intégrité n'est pas garantie. Il s'agit principalement de fichiers dont l'extension n'est pas conforme ou de formats caducs. En filtrant les résultats de la colonne « EXTENSION_MISMATCH » de DROID, nous avons dénombré 292 fichiers corrompus.

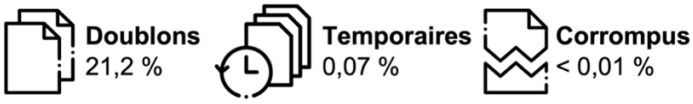

Figure 8 : Synthèse des pourcentages d'occupation du disque par les documents éliminables

(Autrices)

3.4. Répartition des formats

Archifiltre catégorise les fichiers selon 9 types, contre 26 pour DROID. Parmi les types recensés, les données audiovisuelles occupent une part importante. Archifiltre nous permet de faire ressorti les trois principaux groupes : **Image** avec 21 588 éléments, suivi par **Autre** avec 10 911 éléments et **Vidéo** en troisième place avec 747 éléments (Figure 9). La catégorie **Autre** d'Archifiltre a retenu notre attention en raison de sa volumétrie. En effet, le programme semble ne pas avoir été en mesure d'identifier correctement certaines extensions.

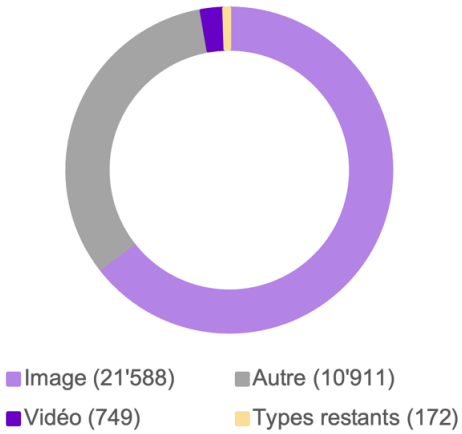

Figure 9 : Répartition par type de formats selon les chiffres d'Archifiltre

(Autrices)

DROID nous offre une identification plus fine des formats et de leurs types. Dans le Comprehensive breakdown à la section File sizes by MIME Type nous trouvons un type (MIME Type) — par exemple **Image** — et un sous-type — une extension de fichier comme .jpeg ou .tiff. Dans la Figure 10, nous avons représenté la répartition des sous-types pour le MIME Type = Image composée de 25 461 éléments.

Figure 10 : MIME Type pour « Image » selon les chiffres de DROID

(Autrices)

3.5. Nomenclature

La nomenclature des documents ne suit pas la directive de nomenclature édictée par l'institution et est hétérogène :

- Les fichiers comme les dossiers contiennent des caractères diacritiques, des caractères spéciaux et des mots vides,
- Les noms manquent de clarté,
- Les séparations édictées par le plan de classement au sein du vrac numérique sont absentes.
- Il n'existe pas d'index propre au service communication pour retracer les décisions de nommage. Cela péjore le taux de réussite d'identification des dossiers, notamment dans le cas de demandes externes.

4. Analyse

Plutôt que de nous baser exclusivement sur les exportations de l'un des logiciels, au risque de perdre ce qui fait leurs forces respectives, nous avons donc opté pour un récolement des données exportées au format .csv. Nous avons affiné manuellement l'identification des formats en nous concentrant particulièrement sur les formats non identifiés ou mal catégorisés. Finalement, nous avons ajouté des catégories utiles à l'analyse des problématiques spécifiques au fond : l'identification des formats propriétaires, et la détection des fichiers représentant un problème pour la conservation.

4.1. État du disque

Dans le tableau comparatif ci-dessous, nous avons reporté les résultats obtenus avec les logiciels en comparaison avec ceux obtenus de notre fichier croisé, une fois les données récolées et nettoyées des doublons (dédoublonnage systématique), fichiers corrompus et temporaires. Nous obtenons un nombre total de 21 306 éléments et de 443 dossiers, pour un volume total de 2,34 Go.

La réduction importante du nombre de dossiers s'explique par la présence de dossiers vides et de nombreux dossiers redondants. Ces derniers sont le résultat d'une imbrication complexe de dossiers — générés automatiquement par les caméras ou les microphones utilisés et inutiles pour le service — et de la duplication de dossiers au sein des environnements de travail de la Suite Adobe.

Tableau 2 — Tableau comparatif de l'état du vrac numérique entre Archifiltre, DROID et de nos données croisées

https://doi.org/10.55790/journals/ressi.2024.e1522

	Archifiltre	DROID	Données croisées
Volume (To)	2,6	2,57	2,34
Dossiers (nb)	832	938	443
Fichiers (nb)	33'418	36'095	21'306

Tableau 2 : Tableau comparatif de l'état du vrac numérique entre Archifiltre, DROID et de nos données croisées

(Autrices)

Dans le vrac nettoyé, nous constatons que les dossiers restants contiennent en moyenne 48 éléments uniques. Ce nombre est trop élevé et nuit à la navigation au sein des documents.

La première date extrême relevée par les logiciels nous a paru étrange — 31. 12 .1979 pour Archifiltre et 1980 pour DROID. Après une rapide recherche en ligne avec les informations à disposition (nom de l'évènement, nom du compositeur et la date), nous en avons déduit qu'il s'agissait de la captation numérisée d'un concert du compositeur à cette date. Ainsi, cette date surprenante pour un vrac exclusivement composé de données nées numériques s'explique par la mauvaise interprétation des métadonnées attachées au fichier. Cette hypothèse est confirmée par le guide d'utilisateur de DROID. Celui-ci stipule que les données contenues dans un fichier peuvent être plus anciennes que le fichier lui-même — dans le cas où un fichier a été copié, ou simplement dactylographié manuellement d'un contenu plus ancien[17]. Une autre explication serait que le réglage de l'horloge de l'ordinateur n'était pas à jour lors de la création, du dépôt ou de la modification (The National Archives, 2023, p. 11). La datation de certains fichiers par les outils d'analyse automatisée est à prendre avec un grain de sel, surtout en l'absence de contexte autour des données. Cependant, nous pouvons conclure que la création du vrac est antérieure aux différentes politiques d'archivage numérique édictées par l'institution.

Les fichiers les plus anciens et les plus récents se situent dans le même répertoire, qui est également le plus volumineux avec 538,43 Go. Nous constatons que 49 % des éléments du vrac numérique ont été modifiés pour la dernière fois entre 2013 et 2015. Finalement, les dates obtenues avec les logiciels ne nous permettent pas de tirer de conclusion définitive sur la fréquence de dépôts effectués.

4.2. Arborescence

La grande profondeur de dossiers sur 13 niveaux nuit à la navigation et à l'identification des documents pertinents. De plus, nous ne retrouvons pas la structure du plan de classement édicté par l'institution. Une brève analyse micro des 109 documents bureautiques nous laissent penser qu'aucun document relatif à la stratégie du service de communication n'est disponible sur le vrac. L'addition de ces éléments nous confirme que le cold storage est utilisé à des fins de déstockage du serveur courant. Il ne permet pas de « témoigner de l'ensemble des activités de [l'institution] » (Lacombe, 2012, p. 36).

[17] « The content of a file (the data within it) may actually be older than the file itself – if a file was copied, or simply typed up manually from an older piece of content" (The National Archives, 2023, p. 11).

4.3. Diagnostic des données

Les fichiers temporaires ont fait l'objet d'une réflexion approfondie, avant d'être étiquetés comme tels. Les informations concernant la conservation ou la suppression de ce type de fichier ont été compliquées à rassembler. Cependant, nous en avons retenu qu'une fois sortis du contexte de création d'un document ou d'une session personnalisée, ces fichiers n'ont plus de valeur d'usage. Ils sont donc à éliminer.

La forte présence des fichiers images dans les doublons (Figure 7) peut s'expliquer en partie par la création de documents habituels des services communication dans un environnement comme la Suite Adobe Creative, et plus particulièrement lors de l'utilisation du logiciel InDesign[18]. Les logiciels de cette suite exigent la création d'un espace de travail, basé sur des liens vers les documents utilisés. Si les images sont déplacées ou renommées, les liens vers les sources sont perdus et doivent être recréés. D'où l'utilisation de doublons de dossiers nécessaires aux divers projets. De manière générale, cela confirme que les fichiers images ne font pas l'objet d'un tri préalable avant leur dépôt sur le cold storage. De plus, la présence de nombreux fichiers « bruts » — le plus souvent fournis par les prestataires de captation — vient augmenter considérablement la taille du vrac. Ces éléments compliquent la prise en charge des données stockées dans un workflow déjà tendu, ce qui entretient un cercle vicieux.

4.4. Répartition des formats

Archifiltre identifie 9 types de fichiers, avec un nombre trop important de types non identifiés dans Autre, et DROID en dénombre 26, trop précis au niveau des formats. Nous avons établi une nouvelle liste de 7 types axés sur les usages du service communication. Notre volonté était d'illustrer les tendances de leur production documentaire et d'identifier les problématiques liées. Cette étape nous a permis de réduire au maximum la catégorie « Autre » et d'attribuer autant de fichiers que possible à un type, sur la base des informations fournies par PRONOM.

Nous constatons que deux types de formats se distinguent fortement : les vidéos et les images. L'ensemble de nos chiffres est répertorié dans le tableau par volumétrie décroissante ci-dessous.

Tableau 3 — Catégories des types de fichiers avec leur poids et leur nombre

[18] Les logiciels de cette suite exigent la création d'un espace de travail, basé sur des liens vers les documents utilisés. Si les images sont déplacées ou renommées, les liens vers les sources sont perdus et doivent être recréés. D'où l'utilisation de doublons de dossiers nécessaires aux divers projets.

	Poids en Go	Nb de fichiers
Vidéo	1 720,300	707
Image	276,900	15 949
Audio	13,287	56
Zip	7,499	48
Texte	0,027	4 429
Bureautique	0,023	109
Autres	0,004	8

Tableau 3 : Catégories des types de fichiers avec leur poids et leur nombre

(Autrices)

Ces chiffres nous montrent que les **vidéos** occupent environ 1,7 To sur le volume total du vrac de 2,34 To, soit près de la moitié. Cependant, ces mêmes fichiers comptabilisent seulement 707 éléments sur un total de 21 306, soit à peine 3 % du nombre total de fichiers. Les **images** occupent environ 277 Go pour 75 % du nombre d'éléments présents sur le vrac (15 949 sur 21 306). Les images ont déjà subi un tri à la suite du dédoublonnage effectué avec les empreintes hashs. Elles sont passées à 15 949 au lieu de 21 306 éléments au début. En dehors des images et des vidéos, les autres types de formats ne sont pas très présents au sein de ce vrac, tant en termes de volume que de nombre d'éléments (liste ci-dessus). Les fichiers textes sont nombreux, mais très légers (inférieurs à 0,1 Go), à l'inverse des fichiers audio, peu nombreux, mais bien plus volumineux (plus de 13 Go).

Au sein de notre nombre restreint de types, nous avons comptabilisé 40 formats différents, que nous avons ensuite distingué en formats ouverts et propriétaires (Figure 11). Cette distinction a été effectuée grâce à l'étude des PIUD[19] fournis par DROID.

Figure 11 — Nombre de types et nombre de formats présents dans le vrac numérique

[19] PIUD pour PRONOM Unique Identifier. Nous avons utilisé les informations de la base de données PRONOM mais également les normes ISO et pages Wikipédia de certains formats afin de confirmer les catégorisations. Notons qu'un PIUD ne change jamais bien que de nouveaux puissent être définis (The National Archives, 2023, p. 12 13).

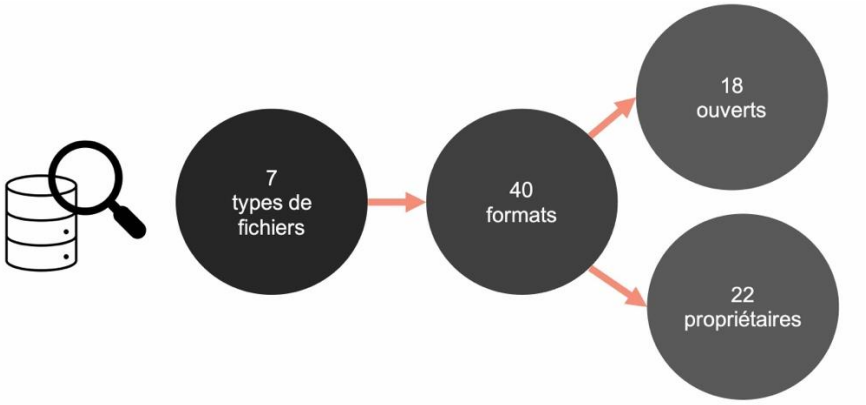

Figure 11 : Nombre de types et nombre de formats présents dans le vrac numérique

(Autrices)

Le format ouvert est « légalement exempté de droits d'utilisation » (TGE-Adonis et al., 2011, p. 11). De par sa transparence, ce format est interopérable, car « il peut être créé, lu et modifié par tous les logiciels destinés à traiter le type du fichier (image, texte, audio, etc.) » (Urfist Méditerranée & Inist-CNRS, 2023).

À l'inverse, le format propriétaire est « contrôlable par une personne ou une entité juridique » (TGE-Adonis et al., 2011, p. 12), ce qui signifie que son encodage est fermé et peut être restreint. Le plus souvent, il nécessite l'utilisation de logiciels spécifiques développés par des entreprises privées. C'est le cas par exemple des formats de l'Adobe Creative Suite, comme le format .psd.

L'identification des formats propriétaires est capitale parce qu'ils sont instables dans le temps, sans garantie de lisibilité ou d'accès dans le futur. Cependant, certains formats propriétaires sont standardisés et préconisés pour la conservation comme c'est le cas des PDF/A qui a fait l'objet d'une norme ISO.

Pour visualiser les enjeux liés à ces formats et distinguer s'ils représentent un problème, nous avons ajouté deux nouvelles catégories binaires distinctes : Propriétaire et Problématique.

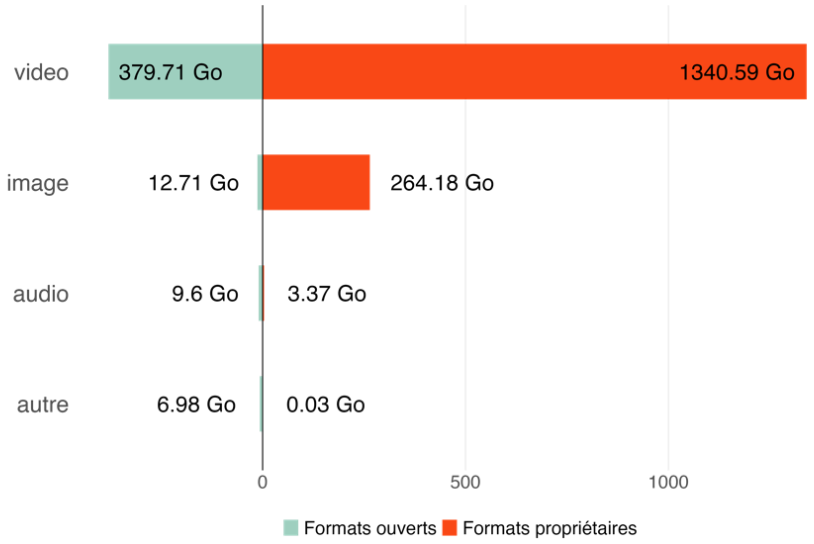

Figure 12 : Occupation du disque par type de fichiers

(Autrices)

La majorité des fichiers vidéo sont propriétaires. Cela s'explique par la part importante de rushs vidéo provenant de captations et ne faisant pas l'objet de conversions systématiques. En parallèle, nous retrouvons des formats. mkv ouverts, propres à la diffusion et à la conservation. Ce pattern se reproduit dans le cas des images. La grande majorité est composée de formats propriétaires comme le .psd ou le .cr2 — un format utilisé par Canon pour sauvegarder des éléments RAW, c'est-à-dire des photographies non traitées de bonne qualité. Nous retrouvons également quelques formats ouverts habituels tels .jpeg, .gif, etc., en moindres proportions.

4.5. Nomenclature

Nous avons constaté que les règles de nommage édictées par l'institution ne sont pas totalement respectées. La nomenclature actuelle suit les besoins à chaud des membres de l'équipe communication, de sorte qu'elle s'inscrit dans la connaissance tacite de l'équipe. La navigation du cold storage est fastidieuse de par les éléments suivants : le nombre élevé de dossiers imbriqués sans réelle fonction, le nom des fichiers ambigus et la nomenclature non systématique — parfois par année, par thème ou encore par auteur·rice.

Tous ces problèmes peuvent être résolus en combinant le respect des directives de nommage de l'institution et en formalisant par écrit les décisions de nomenclature spécifique au service de communication.

Malgré ces problèmes, la longueur maximale des chemins d'accès à un fichier, selon les directives de l'institution, est respectée.

4.6. Éléments complémentaires

Le cold storage regroupe des dossiers et des documents d'une grande variété de contenus : logos, captations vidéo d'évènements — concerts, conférences, etc. —, trailers ou encore rushs. Ces éléments documentent les activités de l'institution et de ses membres. Ils font partie de ce que le service communication souhaite promouvoir et valoriser dans le futur.

Nous avons identifié un certain nombre d'éléments qui ne découlent pas de la production du service communication. Ils appartiennent à d'autres services de l'institution, versés dans le

cold storage de la communication par manque de place sur leur propre serveur. Ces versements ont été effectués sans la consultation du responsable du service communication. Ces documents doivent être versés aux producteur·rice·s d'origine et supprimés définitivement du cold storage du service de communication.

5. Problèmes rencontrés

Durant cette recherche, nous avons rencontré des problèmes utiles à retenir comme points d'attention pour d'autres recherches.

Initialement, nous voulions compléter notre analyse avec un troisième programme spécifique à la validation et la préservation des fichiers numériques audiovisuels : JHOVE. Cependant, la procédure d'installation du programme open source a été extrêmement contraignante et nous n'avons pas été en mesure de l'utiliser.

L'extraction des données a été chronophage et fastidieuse. Nous nous sommes retrouvées face à de multiples échecs d'analyse dus à deux facteurs dont nous n'avions pas connaissance au préalable. Tout d'abord, la génération de fichiers temporaires est conséquente et nécessaire à l'analyse automatisée. Celle-ci a plusieurs fois saturé le disque de l'ordinateur mis à disposition par le service, entraînant un échec d'export. Le deuxième est dû à un type de fichier vidéo propriétaire (.cr2) volumineux qui a représenté un challenge aux deux logiciels.

Le récolement des données a été effectué du mieux que nous avons pu avec les données récoltées. Nous n'avons pas pu récupérer les hashs depuis DROID, et les chemins d'accès aux fichiers ne sont pas formatés de la même manière entre les deux logiciels. Pour ces raisons, nous avons dû procéder à un récolement par nom de fichier et par poids du document. Cette étape a pu être à l'origine de disparités ou d'erreurs dans les nombres retenus pour l'analyse.

Concernant les dates de dernière modification, l'interface d'Archifiltre peut prêter à confusion. En effet, les dates à prendre en compte sont celles indiquées dans l'onglet « Général » du programme, et non celles indiquées dans l'outil de recherche. Dans ce dernier cas, la date de dernière modification correspond à la date de consultation des données sur le logiciel et non pas à la date de dernière modification du fichier analysé.

6. Conclusion

Trois objectifs ont guidé la réalisation de ce mandat de recherche : identifier le contenu stocké dans le cold storage et fournir des points de traitement des données, afin de faire ressortir les données nécessitant une évaluation plus fine pour la promotion et la valorisation des activités de l'institution. Après un premier nettoyage suggéré des données, nous avons identifié que le plus grand challenge de l'équipe réside dans les formats propriétaires, et plus particulièrement les vidéos et les images. En effet, ces formats sont les plus nombreux, les plus lourds et sont hautement instables dans le temps.

À l'issue de ce projet, nous avons identifié trois types d'axes d'amélioration : les pratiques du service, la gestion du volume des données ainsi que quelques critères utiles à la valorisation future des données. Dans une perspective d'archivage à long terme, nous recommandons une approche d'archivage dès le début de la création documentaire, selon le concept d'archiving

https://doi.org/10.55790/journals/ressi.2024.e1522

by design (partie 1.3.9), afin d'alléger la charge de l'équipe et de faciliter l'accessibilité à l'information.

Le service communication doit intégrer le plan de classement à son workflow ainsi que les principes directeurs de l'institution (partie 2.2.3). Conformément à la législation, le service est responsable des données qu'il produit, ce qui implique les principes d'intégrité, de fiabilité, et de sécurité. Il doit contextualiser sa production, ce qui passe par l'intégration des documents administratifs officiels et authentiques relatifs à sa stratégie et ses missions, ainsi que par l'implémentation de métadonnées. L'équipe doit également être au fait des durées de conservations des documents produits afin de ne pas conserver des données personnelles ou sensibles au-delà des délais prévus par la loi. Tout document ne découlant pas de ses activités est à retourner aux producteur·rice·s d'origine et à supprimer définitivement des serveurs de la communication. Une nomenclature et des processus de travail homogénéisés facilitent fortement l'identification des documents importants. Finalement, nous recommandons à l'équipe d'intégrer dans son workflow un temps dédié au traitement des données afin de ne pas être débordée par un vrac numérique et d'être en conformité avec les diverses lois, normes et directives.

Afin de ne pas être à nouveau confrontée à un vrac numérique, l'équipe doit distinguer les documents de travail des documents définitifs. Au cours de notre recherche, nous avons constaté que l'équipe a une connaissance tacite du versionnage des documents, et qu'elle conserve toutes les versions « au cas où ». Cependant, nos échanges et notre analyse confirment que l'équipe ne revient pas sur les documents d'un projet une fois celui-ci clos. Ainsi, la période de clôture pourrait être un moment propice au tri des documents de travail et à la conservation conforme des documents définitifs. Ce principe permettrait de :

- Éliminer les rushs et autres fichiers sources de prestataires externes à la faveur de la documentation produite par le service, et
- Éliminer les diverses versions de travail à la faveur du document définitif produit (trailer, affiche, flyer, newsletter, etc.).

Concernant la valorisation patrimoniale, nous recommandons d'établir une liste de critères propres aux contenus produits par le service communication selon trois catégories : esthétiques — qualité du support et de l'image —, contenu — rareté du sujet, technique, documentation de l'activité —, et historiques. Cette dernière catégorie se réfère aux documents relatifs à des périodes particulières afin de documenter non seulement l'institution, mais également la société. Deux exemples sont la création de l'institution ou encore l'activité de celle-ci durant la pandémie de covid-19.

Conformément aux recommandations de la National Archives, nous recommandons vivement la mise en place d'un guide pratique et une formation du personnel sur le choix des éléments à conserver et l'intégration du tri dans les processus de travail (The National Archives, 2013).

Finalement, les deux logiciels sont complémentaires dans l'appréhension d'un vrac de données audiovisuelles. Ils offrent respectivement une approche macro et micro sans laquelle nous n'aurions pas pu avoir une compréhension holistique. Ainsi, Archifiltre a permis de visualiser la structure et de naviguer dans l'arborescence, et DROID de profiler les données, étape essentielle à l'évaluation de leur santé. Cependant, il s'agit de logiciels d'analyse et non

de traitement des données[20]. À l'heure actuelle, nous ne pouvons que recommander l'utilisation combinée des logiciels pour l'analyse d'un vrac audiovisuel, mais le traitement des données nécessitera un logiciel supplémentaire, dédié au traitement, l'aide de professionnel·le·s de l'information et l'intervention du service informatique. Tant paramètres à considérer lors de la prise en main de vracs numériques.

Bibliographie

Aas. (2018). Developing the « archive it » button. Arbido, 2018/2. https://arbido.ch/en/ausgaben-artikel/2018-1/automatisierung-versprechen-oder-drohung/developing-the-archive-it-button

AFS, A. fédérales suisses. (2022, octobre 31). Valeur archivistique. https://www.bar.admin.ch/bar/fr/home/informationsmanagement/archivwuerdigkeit.html

Archifiltre, F. des M. sociaux. (2023). Archifiltre. archifiltre.fabrique.social.gouv. https://archifiltre.fabrique.social.gouv.fr/

Association des archivistes français. (2018). Fiche pratique AMAE n°22 : Réflexion sur le vrac numérique. https://www.archivistes.org/IMG/pdf/fp22_reflexions_vrac_numerique.pdf?7512/9f51d700023%209d6b41b3be59a7b0f2c3748adfeae

Atlan. (2023, août 3). Data Lake vs Data Swamp : Differences & Cautionary Steps. Atlan.Com. https://atlan.com/data-lake-vs-data-swamp/

Bardiot, C. (2022). Comment transmettre l'héritage des arts de la scène ? In Les devenirs numériques des patrimoines (p. 304 315). Maison des sciences de l'homme.

Bavaud, A., Bischoff, S., & Bussard, D. (2022). Automatisation des fonctions archivistiques pour les données textuelles : Quels outils et quelles fonctionnalités pour l'archiviste ? [Haute École de Gestion de Genève (HEG-GE)]. https://sonar.rero.ch/hesso/documents/322791

Belovari, S. (2019). Expedited Digital Appraisal for Regular Archivists: An MPLP-type Approach. Journal of Archival Organization, 14(1 2), 55 77. https://doi.org/10.1080/15332748.2018.1503014

Bennani, R., Hategekimana, A., & Rey Rodriguez, A. (2022). Automatisation des fonctions archivistiques pour les données non textuelles : Le cas des photographies en Suisse [Haute École de Gestion de Genève (HEG-GE)]. https://sonar.rero.ch/hesso/documents/322358

Bischoff, S. (2022). État des lieux des pratiques d'évaluation aux Archives de l'État de Neuchâtel et exploration de leur traduction en fonctionnalités pour le logiciel ArchiSelect. https://sonar.ch/global/documents/322898

[20] Archifiltre tente cependant le pas en proposant certaines prestations facilitant un traitement a posteriori, comme par exemple le dédoublonnage systématique, la restructuration de l'arborescence ou la suppression de documents, le tout exportable dans un script de suppression.

Bunn, J. (2023). AI for Appraisal and Selection: A personal reflection. Arbido, 2023/1. https://www.arbido.ch/de/ausgaben-artikel/2023/archiv-der-zukunft/ai-for-appraisal-and-selection-a-personal-reflection

Bussard, D. (2022). Outils et méthodologie pour le tri archivistique de supports de données complexes [Haute École de Gestion de Genève (HEG-GE)]. https://sonar.rero.ch/global/documents/322896

Conférence des recteurs et des principaux des universités du Québec. (2009). Mesures transitoires et bonnes pratiques de gestion des documents numériques. https://www.bci-qc.ca/wp-content/uploads/2023/03/mesures_transitoires_bonnes_pratiques_GDN-10juin09_avis.pdf

Coutaz, G. (2016a). Archives en Suisse (Presses polytechniques et universitaires romandes).

Coutaz, G. (2016b). La croissance et la maîtrise des masses documentaires. arbido.ch. https://arbido.ch/fr/edition-article/2016/détruire-pour-conserver/la-croissance-et-la-maîtrise-des-masses-documentaires

Couture, C. (1996). L'évaluation des archives. État de la question. Archives, 28(1), 3 31.

Couture, C. (2011). Les fonctions de l'archivistique contemporaine. Presses de l'Université du Québec.

Data swamp: Definition, challenges on a data lake, solutions. (2023, juin 13). Starburst.Io. https://www.starburst.io/learn/data-fundamentals/data-swamp/

EAG - European Archives Group. (2023). Archiving by Design Whitepaper. https://commission.europa.eu/system/files/2023-06/Whitepaper%20AbD_en.pdf

File Extension. (2023). file-extension.info. https://www.file-extension.info/fr

Gagnon-Arguin, L. (2003). La création. In Les fonctions de l'archivistiques contemporaine, COUTURE, Carol (Québec : Presse de l'Université du Québec, p. 69 102).

Grand dictionnaire terminologique. (2023). Données non structurées. In Grand dictionnaire terminologique de l'Office québécois de la langue française (en ligne). https://vitrinelinguistique.oqlf.gouv.qc.ca/fiche-gdt/fiche/8873662/donnees-non-structurees

Greene, M. A. (2010). MPLP : It's Not Just for Processing Anymore. The American Archivist, 73(1), 175 203.

Greene, M., & Meissner, D. (2005). More Product, Less Process: Revamping Traditional Archival Processing. The American Archivist, 68(2), 208 263. https://doi.org/10.17723/aarc.68.2.c741823776k65863

Hoolt, V. (2023). Archiving by design: Theory and practice in the Netherlands. Arbido, 2023/1. https://arbido.ch/de/ausgaben-artikel/2023/archiv-der-zukunft/archiving-by-design-theory-and-practice-in-the-netherlands

International Standard Organisation. (2012). ISO 14721:2012, Systèmes de transfert des informations et données spatiales—Système ouvert d'archivage d'information (SOAI)—Modèle de référence. https://www.iso.org/fr/standard/57284.html

International Standard Organisation. (2016). ISO 15489-1:2016, Information et documentation—Gestion des documents d'activité—Partie 1 : Concepts et principes. https://www.iso.org/fr/standard/62542.html

International Standard Organisation. (2020). ISO 30300:2020, Information et documentation—Systèmes de gestion des documents d'activité—Principes essentiels et vocabulaire. https://www.iso.org/fr/standard/74291.html

International Standard Organisation. (2022a). ISO 13008:2022, Information et documentation—Processus de conversion et migration des documents d'activité numériques. https://www.iso.org/fr/standard/75569.html

International Standard Organisation. (2022b). ISO 24143:2022, Information and documentation—Information Governance—Concept and principles. https://viewer.snv.ch/product/28239?langUI=fr&filePath=66e976dd-5ef9-40af-8a44-a57b376ff1e5.pdf&fileType=Pdf

International Standard Organisation. (2022c). ISO/TR 26122:2008, Information et documentation—Analyse des processus pour la gestion des informations et documents d'activité. https://www.iso.org/fr/standard/43391.html

InterPARES Trust AI. (2023). Terminology Database—Information asset [English]. Interparestrustai.Org. https://interparestrustai.org/terminology/term/information%20asset

Koch, R. (2018). DRAINING THE DATA SWAMP: An organization's data lake is only as good as the preparation and maintenance planning that go into creating it. Strategic Finance, 99(12), 62 64.

Lacombe, C. (2012). Les principes directeurs de l'évaluation archivistique en question. Archives, 44(1), 35 43.

L'Assemblée fédérale de la Confédération suisse. (2020). Loi fédérale du 25 septembre 2020 sur la protection des données (nLPD). https://www.fedlex.admin.ch/eli/oc/2022/491/fr

Lefort, L. (2018). Valorisation des photographies contenues dans les archives de la justice et de la police fribourgeoises : Proposition d'une procédure de sélection, de description et de diffusion par les nouveaux médias. https://sonar.ch/global/documents/314861

Lenartz, S. (2020). Digital ist besser?. Möglichkeiten der automatisierten Aufbereitung und Bewertung von Fileablagen mit Python am Beispiel einer digitalen Fotosammlung. https://d-nb.info/1212867793/34

lerobert.com. (2023). Vrac. In Le Robert Dico en ligne. https://dictionnaire.lerobert.com/definition/vrac

Makhlouf Shabou, B. (2023a). Archivistique à l'ère de l'IA : Opportunités, défis et besoin d'utilisation responsable. arbido, 2023/1. https://arbido.ch/fr/edition-article/2023/archives-du-futur/archivistique-a-lere-de-lia-opportunites-defis-et-besoin-dutilisation-responsable

Makhlouf Shabou, B. (2023b, 2024). Gouvernance des données.

Makhlouf Shabou, B., Tièche, J., Knafou, J., & Gaudinat, A. (2020). Algorithmic Methods to Explore the Automation of the Appraisal of Structured and Unstructured Digital Data. Records Management Journal, 30(2), 175 200. https://doi.org/10.1108/RMJ-09-2019-0049

Manaher, S. (2021, août 24). Automation Vs. Automatization, Differences & Uses Of Each One [Thecontentauthority.com]. https://thecontentauthority.com/blog/automation-vs-automatization

Memoriav. (2023). Notre Mission. memoriav.ch. https://memoriav.ch/fr/mission/

Merriam Webster. (2023, novembre 9). Automation. Merriam-Webster.Com. https://www.merriam-webster.com/dictionary/automation

Mersiovsky, K. (2014, décembre 14). The Pros and Cons of “MPLP”. Archives & Memory. https://medium.com/archives-records/the-pros-and-cons-of-mplp-586b0efc8fba

Office québécois de la langue française. (2023). Empreinte numérique. oqlf.gouv.qc.ca. http://www.oqlf.gouv.qc.ca/RESSOURCES/bibliotheque/dictionnaires/terminologie_sec_informatique/empreinte_numerique.html

Oguey, G., & Schneiter, P. (2018). ArchiSelect, ou quand l’évaluation s’automatise. arbido, 2018/2. https://arbido.ch/fr/edition-article/2018/automatisierung-versprechen-oder-drohung/archiselect-ou-quand-lévaluation-sautomatise

Organisation Internationale de Normalisation. (2016). Information et documentation Gestion des documents d’activité Partie 1 : Concepts et principes (SN ISO 15489-1:2016; Numéro SN ISO 15489-1:2016). International Organization for Standardization. https://viewer.snv.ch/product/261553/fr

Organisation Internationale de Normalisation. (2018). Electronic document management Design and operation of an information system for the preservation of electronic documents—Specifications (ISO 14641:2018; Numéro ISO 14641:2018). International Organization for Standardization. https://viewer.snv.ch/product/196952?langUI=fr&filePath=35293bf5-034f-4c9d-a8c0-1543b0f84780.pdf&fileType=Pdf

Organisation Internationale de Normalisation. (2022a). Data quality Part 2: Vocabulary (ISO 8000-2:2022; Numéro ISO 8000-2:2022). International Organization for Standardization. https://viewer.snv.ch/product/829095/fr

Organisation Internationale de Normalisation. (2022b). Information and documentation—Information Governance—Concept and principles (ISO/FDIS 24143:2022; Numéro ISO/FDIS 24143:2022). International Organization for Standardization. https://viewer.snv.ch/product/28239?langUI=en&filePath=66e976dd-5ef9-40af-8a44-a57b376ff1e5.pdf&fileType=Pdf

Phillips, J. (2015). A Defense of Preservation in the Age of MPLP. The American Archivist, 78(2), 470 487. https://doi.org/10.17723/0360-9081.78.2.470

République et canton de Genève. (2000a). Loi sur les archives publiques (LArch). https://ge.ch/archives/media/site_archives/files/imce/pdf/lois/20210412_larch.pdf

République et canton de Genève. (2000b). Loi sur les archives publiques (LArch). https://ge.ch/archives/media/site_archives/files/imce/pdf/lois/20210412_larch.pdf

République et canton de Genève. (2021). Loi sur l’information du public, l’accès aux documents et la protection des données personnelles (2) (LIPAD). https://silgeneve.ch/legis/index.aspx

Scarpulla, M. (2016). Archives, danse et recréation. Une introduction. Archives, 46(1), 15 34. https://doi.org/10.7202/1035720ar

Texier, B. (2022a). Le Data Lab de la BNF à l'assaut des données inexplorées. Archimag, 351, 27 28.

Texier, B. (2021a, février 3). Traiter un vrac numérique : Mode d'emploi | Archimag. archimag.com. https://www.archimag.com/demat-cloud/2021/05/05/traiter-vrac-numerique-mode-emploi

Texier, B. (2021b, mars). Vrac numérique : Comment mettre de l'ordre, Numéro 342. archimag.com. https://www.archimag.com/le-kiosque/mensuel-archimag/mag-342/vrac-numerique-mettre-ordre/PDF

Texier, B. (2022b, février 3). Vrac numérique : Un chantier qui ne s'improvise pas. archimag.com. https://www.archimag.com/demat-cloud/2021/05/05/vrac-numerique-chantier-improvise-pas

TGE-Adonis, SIAF, S. interministériel des A. de F., & CNRS. (2011). Guide méthodologique pour le choix de formats numériques pérennes dans un contexte de contexte de données orales et visuelles (ADONIS/SIAF/CINES-GM-0.5). https://francearchives.gouv.fr/file/9c2af48c1e112ed745711eb7c7e15e6d70b8c0f9/static_4923.pdf

The File Format Database. (2023). [Fileinfo.com]. https://fileinfo.com/

The National Archives. (2013). Best practice guide to appraising and selecting records for The National Archives. https://cdn.nationalarchives.gov.uk/documents/information-management/best-practice-guide-appraising-and-selecting.pdf

The National Archives. (2022). What is appraisal? https://cdn.nationalarchives.gov.uk/documents/information-management/what-is-appraisal.pdf

The National Archives. (2023). DROID: User guide. https://cdn.nationalarchives.gov.uk/documents/information-management/droid-user-guide.pdf

The Source for File Extension Information. (2023). [File-extensions.org]. https://www.file-extensions.org/

Urfist Méditerranée, & Inist-CNRS. (2023, août 3). Format ouvert ou fermé ? doranum.fr. https://doranum.fr/stockage-archivage/quiz-format-ouvert-ou-ferme_10_13143_mcwq-qs64/