h e g

Haute école de gestion
Genève

Automatisation des fonctions archivistiques pour les données textuelles : quels outils et quelles fonctionnalités pour l'archiviste ?

Mémoire de Recherche réalisé

par :

Aurélie BAVAUD

Sébastien BISCHOFF

Denis BUSSARD

sous la direction de :

Basma MAKHLOUF-SHABOU, professeure HES

Genève, 14 janvier 2022

Haute École de Gestion de Genève (HEG-GE)

Master en Sciences de l'Information

Hes·so GENÈVE
Haute Ecole Spécialisée de Suisse occidentale

Déclaration

« Nous attestons avoir réalisé seul·e·s le présent travail, sans avoir utilisé des sources autres que celles citées dans la bibliographie. »

Fait à Genève, le 14 janvier 2022

Aurélie Bavaud

Sébastien Bischoff

Denis Bussard

Résumé

La surabondance informationnelle, accentuée encore par l'avènement du numérique, pose des problèmes spécifiques en termes de traitement et d'accès aux documents. Pour répondre à ces défis, et face à la menace de *black boxes*, *data swamps* et autres risques informationnels menaçants les services d'archives et les sociétés qui reposent sur celles-ci, les archivistes doivent adapter leurs pratiques et exploiter les nouvelles technologies.

Annoncée partout et promise depuis des années, l'automatisation des fonctions archivistiques se fait encore attendre. Face à un champ de recherche aussi vaste, hétérogène et technique, les praticien·ne·s ont de quoi être désemparé·e·s. Le concept d'automatisation reste difficile à définir et son champ d'application demeure flou.

Nous proposons, dans ce travail, de faire le point sur ce domaine d'étude relativement nouveau, qu'est la *Computational Archival Science*. Par une étude documentaire basée sur la littérature scientifique et professionnelle, nous proposons un panorama des outils actuellement disponibles pour la gestion des archives électroniques et tentons de cerner au plus près les possibilités offertes par l'automatisation des fonctions archivistiques pour les données textuelles.

Grâce aux données collectées sur les outils informatiques et via l'extraction et la normalisation des fonctionnalités qu'ils proposent, un instrument d'aide à la décision, sous la forme d'un tableau synoptique et d'un diagramme de membrures, a été créé. Notre ambition est de rendre cet instrument, qui peut être complété et reproduit, utile aussi bien aux développeurs pour cibler leurs efforts sur des tâches peu exploitées, qu'aux utilisateurs, pour choisir les outils qui répondront le mieux à leurs besoins. Afin de mettre à l'épreuve notre méthode d'évaluation, nous avons effectué des études de cas avec trois outils de notre liste.

Table des matières

Déclaration i

Résumé ii

Table des matières iii

Liste des figures v

Liste des abréviations vi

Glossaire vii

1. Introduction 1

1.1 Problématique et contexte 1

1.2 Objectifs et question de recherche 2

1.3 Revue de littérature 3

1.3.1 Archivage numérique et fonctions archivistiques 3

1.3.2 Automatisation des fonctions archivistiques 3

1.3.3 Automatisation grâce à l'intelligence artificielle 4

1.4 Définition des concepts principaux 5

1.4.1 L'automatisation 5

1.4.2 Les fonctions archivistiques 6

1.4.3 Système ouvert d'archivage d'information (OAIS) 8

2. Méthodologie 9

2.1 Typologie de la recherche 9

2.2 Étude documentaire 9

2.2.1 Choix du corpus et collecte des documents 9

2.2.2 Collecte des informations à partir d'un corpus documentaire 10

2.2.3 Traitement des informations : normalisation des fonctionnalités 11

2.2.4 Conception d'une grille d'analyse pour les fonctions archivistiques 12

2.2.5 Analyse et exploitation des résultats 13

2.3 Études de cas 14

2.3.1 Population et échantillon 14

2.3.2 Choix des outils testés 17

3. Résultats 18

3.1 Tableau synoptique 18

3.2 Visualisation 18

3.2.1 Définitions des tâches 19

3.2.2 Définitions des fonctionnalités 23

3.2.3 Un outil pour le développeur....27

3.2.4 Un outil pour l'utilisateur 30

3.3 Études de cas 32

3.3.1 Étude de cas n° 1 : Archifiltre 32

3.3.2 Étude de cas n° 2 : DROID (Digital Record Object Identification) 38

3.3.3 Étude de cas n°3 : Karen's Directory Printer 43

4. Discussion 47

4.1 Discussion des résultats 47

4.1.1 Du soutien à la réalisation d'une tâche archivistique 47

4.1.2 Des tâches techniques, intellectuelles et humaines....48

4.1.3 De l'utilité de certaines fonctionnalités : le cas du *NLP*....49

4.1.4 Des outils provenant d'horizons très divers 50

4.2 Discussion de la méthode 51

4.2.1 Exhaustivité non garantie des outils et des fonctionnalités 51

4.2.2 Disparité des informations mises à disposition 52

4.2.3 Granularité inégale des fonctionnalités....53

5. Conclusion 55

Bibliographie 58

Annexe 1 : Tableau synoptique des outils 66

Annexe 2 : Tableau synoptique des projets et autres références 87

Annexe 3 : Liaisons Fonctionnalités-Tâches (tableau à double entrée) . 106

Annexe 4 : Diagramme 107

Liste des figures

Figure 1 : Situation intenable pour la gestion de la masse documentaire 2

Figure 2 : Entités fonctionnelles OAIS 8

Figure 3 : Extrait du tableau synoptique. 18

Figure 4 : Représentation graphique des liens entre *tâches* et *fonctionnalités*.....28

Figure 5 : La tâche *Indexer*.....29

Figure 6 : Les fonctionnalités utilisant le « NLP » 29

Figure 7 : Archifiltre 30

Figure 8 : DROID 31

Figure 9 : Karen's Directory Printer.....31

Figure 10 : Visualisation d'arborescence « en stalactite » dans Archifiltre 35

Figure 11 : Interface Dédoublonnage dans Archifiltre 37

Figure 12 : Interface DROID.....40

Figure 13 : Rapport synthétique DROID 41

Figure 14 : Dédoublonnage dans DROID 42

Figure 15 : Interface principale de Karen's Directory Printer 45

Figure 16 : Personnalisation de la recherche dans Karen's Directory Printer.....45

Liste des abréviations

AIP *Archival Information Package*

DIP *Dissemination Information Package*

GED Gestion électronique des Documents

LOD *Linked Open Data*

MD5 *Message Digest* 5

MPLP *More Product Less Process*

MSF Médecins Sans Frontières

NARA *The National Archives and Records Administration*

NLP *Natural Language Processing*

OAIS *Open Archival Information System*

OCR *Optical Character Recognition*

ONG Organisation Non Gouvernementale

PII *Personally Identifiable Information*

RDF *Resource Description Framework*

SAE Système d'archivage électronique

SIP *Submission Information Package*

Glossaire

Automatisation « Une automatisation est une technique ou un ensemble de techniques ayant pour but de réduire ou de rendre inutile l'intervention d'opérateurs humains dans un processus où cette intervention était coutumière. [...] Elle tend donc à économiser l'intervention humaine sous toutes ses formes » (Encyclopædia Universalis, 2022)

Born-digital Né numérique en français, « document créé directement sur ordinateur et qui ne contient que des éléments générés par le logiciel qui l'a créé, que ce soit du texte ou des dessins. On parlera aussi de document numérique natif. » (Banat-Berger et Huc 2011)

Cheksum (somme de contrôle) Séquence de données calculée à partir d'un fichier qui permet de vérifier que l'intégrité du fichier a été préservée lors d'une copie, d'une opération de stockage ou de transmission.

Empreinte numérique « Séquence de caractères alphanumériques de longueur fixe, qui représente le contenu d'un message sans le révéler, dont la valeur unique est produite par un algorithme de hachage. » (Grand dictionnaire terminologique 2012)

Linked Open Data « *Linked Open Data defines a vision of globally accessible and linked data on the internet based on the RDF standards of the semantic web*. » (W3C 2019)

Machine Learning Le *machine learning* est une technologie d'intelligence artificielle basée sur les statistiques qui permet aux ordinateurs d'apprendre sur la base de jeux de données, sans programmation préalable.

Message Digest 5 (MD5) Algorithme de hachage permettant d'obtenir une empreinte numérique d'un fichier.

Métadonnées *Metadata* en anglais, « *Information that characterizes another information resource, especially for purposes of documenting, describing, preserving or managing that resource.* » (INTERPares 2018)

Natural Language Processing (NLP) Le *Natural Language Processing*, ou traitement automatique du langage est une branche de l'intelligence artificielle, du *machine learning* et contient des éléments de linguistique. Son objectif est de « comprendre » et analyser le langage naturel afin d'en extraire des connaissances sans intervention humaine (Chaumartin, Lemberger 2020).

Personally identifiable information (PII) « *1. Data that allows a specific individual to be recognized. – 2. Restricted, private data that can be linked to a specific individual.* » (INTERPares 2018)

1. Introduction

1.1 Problématique et contexte

Nous assistons à une croissance exponentielle de la masse documentaire : si une mutation avait déjà été observée dans le monde papier (Weill 1990), elle a pris une autre dimension avec l'avènement du numérique et l'entrée dans l'ère du *Zettabyte*[1] (Floridi 2010). Cette surabondance est devenue une composante fondamentale de la réflexion archivistique (Coutaz 2016).

À titre d'exemple, en Suisse, l'ensemble des services d'archives cantonaux a, à ce jour, accumulé 385 kilomètres linéaires d'archives papier accessibles et l'accroissement pour la seule année 2020 est de 10 kml (Conférence des directrices et directeurs d'Archives suisses 2021). Pour les archives numériques, les chiffres sont encore plus affolants : au sein des National Archives and Records Administration (États-Unis), un rapport a estimé en 2013 que 95 % des archives numériques étaient non traitées (Trace 2021).

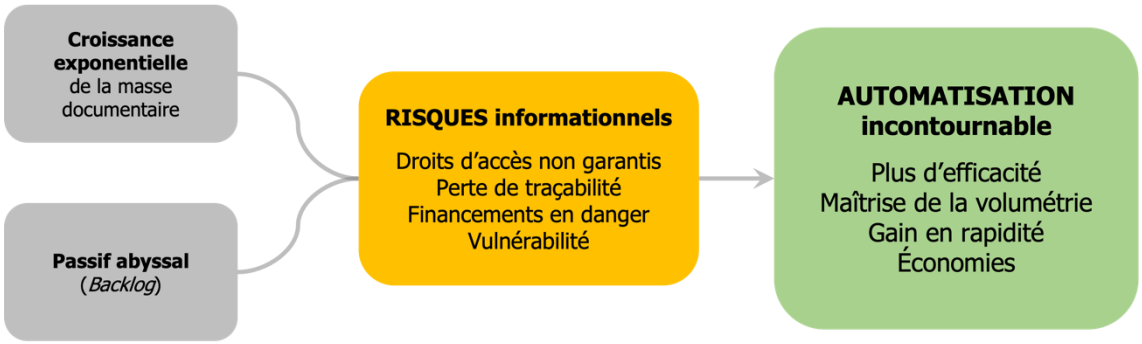
Une telle croissance sans limite s'accompagne inévitablement de pertes (Kecskeméti, Körmendy 2014). Des siècles d'archives non traitées (Greene, Meissner 2005) induisent des risques informationnels réels : nous créons de multiples zones d'ombres sur lesquelles les professionnel·le·s de l'information n'ont aucun contrôle et qui rendent le partage des contenus impossible. Les missions patrimoniales ne sont ainsi pas remplies – dans ces conditions, la valorisation et l'exploitabilité des données sont par exemple impossibles. Ce manque de transparence pose de graves problèmes légaux de droit d'accès aux archives et cette situation pourrait avoir des conséquences sérieuses pour le domaine : les bailleurs de fonds pourraient en venir à questionner le financement des services d'archives (Greene, Meissner 2005). Pour maîtriser ce passif, il faut ainsi une rupture avec les anciennes pratiques (Coutaz 2014).

Si nous changeons bien de paradigme, le support numérique « ne doit avoir aucun impact sur la démarche archivistique qui demeure pertinente et nécessaire » (Béchard, Fuentes Hashimoto, Vasseur 2020). Une automatisation (partielle ou complète) des fonctions archivistiques paraît ainsi être la solution incontournable (Morisod 2018) pour diminuer le passif et permettre aux services d'archives de traiter en flux continu les

[1] L'ère du *Zettabyte* a débuté quand le trafic internet global a dépassé 10^{21} bytes, soit au début des années 2010.

nouvelles productions documentaires avec les ressources humaines limitées à disposition.

Figure 1 : Situation intenable pour la gestion de la masse documentaire



1.2 Objectifs et question de recherche

Jusqu'à ce jour, et dans l'état actuel des connaissances, l'automatisation paraît être au mieux une ritournelle, au pire une chimère. L'automatisation concrète et effective des fonctions archivistiques se fait pourtant encore attendre. Si l'on parle d'automatisation depuis les années 1960 déjà (Bunn 2016), les recherches appliquées dans le domaine archivistique sont plutôt récentes – elles datent approximativement de l'avènement de l'informatique grand public à partir des années 1990.

Ce domaine d'étude est donc relativement nouveau – comme en témoigne la constitution d'un champ d'étude propre à partir de 2016, la *Computational Archival Science* (Marciano et al. 2018 ; Payne 2018). Confondue parfois avec la (préalable) systématisation ou avec l'(omnipotente) intelligence artificielle – représentant toutes deux un extrême du large spectre de l'automatisation –, il est encore difficile d'identifier de manière précise ce que le concept d'« automatisation » recouvre et, à plus forte raison, quel est son champ d'application et quel est son potentiel dans le domaine archivistique.

Face à un domaine de recherche aussi vaste et hétérogène, voire touffu, les praticien·ne·s ont de quoi être désemparés. Ainsi, les **objectifs** visés par notre projet de recherche sont :

- Dresser un état des lieux des initiatives existantes (projets, outils et fonctionnalités) en matière d'automatisation des fonctions archivistiques pour les données textuelles ;
- Identifier quelles fonctions archivistiques sont ou ne sont pas soutenues par lesdites initiatives ;
- Proposer une méthodologie reproductible et généralisée pour répertorier les outils contribuant à l'automatisation ;

- Proposer un instrument d'aide à la décision pour sélectionner l'outil le plus utile à la bonne mise en œuvre d'un traitement documentaire numérique ;
- Développer des études de cas grâce au test de trois outils distincts avec un même jeu de données.

Enfin, notre **question de recherche** est la suivante :

- Quelles sont, à l'heure actuelle, les possibilités offertes par l'automatisation des fonctions archivistiques pour le traitement des données textuelles ?

1.3 Revue de littérature

1.3.1 Archivage numérique et fonctions archivistiques

Dans leur guide intitulé *Les archives électroniques*, Lorène Béchard, Lourdes Fuentes Hashimoto et Édouard Vasseur définissent l'archivage de la manière suivante : « activité qui consiste à gérer et à organiser l'information dans le temps, quel que soit son support, pour la rendre accessible durablement, bien au-delà de la durée de vie des supports » (2020b, p. 6). L'archivage numérique doit ainsi être abordé dans le contexte plus large de l'archivage en général et de ses fonctions traditionnelles.

Ces fonctions ont été définies notamment par Carol Couture dans *Les fonctions de l'archivistique contemporaine* (1999) (Voir le point 1.4.2 Les fonctions archivistiques dans le présent travail). Dans cet ouvrage, l'auteur reste cependant très focalisé sur l'archivage analogique. Couture s'intéresse, avec Marcel Lajeunesse, à la problématique de l'archivage numérique dans une autre monographie parue quinze ans plus tard : *L'archivistique à l'ère du numérique : les éléments fondamentaux de la discipline* (2014). Enfin, Françoise Benat-Berger (2012) dédie un chapitre entier de son volume *Les chantiers du numérique. Dématérialisation des archives et métiers de l'archiviste* aux fonctions archivistiques à l'ère du numérique.

1.3.2 Automatisation des fonctions archivistiques

En 2014, les *National Archives and Records Administration* (NARA) publient un rapport, qui marque un jalon important dans le domaine archivistique, sous le titre : *Managing Government Records Directive – Automated Electronic Records Management Report/Plan*. Les NARA y font le constat de la nécessité de l'automatisation dans la gestion documentaire et proposent quelques pistes d'application ainsi qu'une typologie du spectre de l'automatisation (de « *no automation* » à « *autoclassification* »). D'autres avant eux ont également plaidé dans le même sens, dont Steve Bailey avec un article au titre évocateur : « Forget electronic records management, it's automated records management that we desperately need » (2009).

Les recherches se sont jusqu'à présent surtout concentrées sur l'évaluation des documents, dans l'espoir notamment de réduire la masse des archives numériques conservées (Harvey, Thompson 2010 ; Lee 2018; Makhlouf Shabou et al. 2020). Dépassant certes la question de l'archivage numérique, il convient également de citer les travaux de Greene et Meissner autour du concept de *More Product Less Process* (MPLP) qui ont passablement influencé le champ de la recherche (Greene 2010; Belovari 2017). Penn (2019) en fait d'ailleurs une bonne présentation et l'illustre avec une étude de cas tandis que Godmann (2017) propose des solutions pragmatiques pour accélérer le traitement documentaire. Poursuivant le même objectif, Cyndi Shein expose son flux de travail en détaillant les étapes franchies et les outils utilisés pour le traitement archivistique complet d'une collection de documents *born-digital* hybrides en adoptant une approche de type MPLP (2014).

Les expériences ont jusqu'ici principalement été menées dans le domaine des archives publiques et administratives. En outre, elles ont le plus souvent porté sur un corpus de courriels à traiter grâce à des outils informatiques *ad hoc* – citons notamment les recherches de Alberts et Vellino (2013 ; 2016). Les archives privées ont, en revanche, été relativement peu abordées par la recherche scientifique et professionnelle, mis à part les travaux de Kim et al.(2006) sur les manuscrits d'Arnold Wesker et la très intéressante étude de cas menée par Susanne Belovari à Ludwigsburg en 2017 sur un fonds privé (2017).

1.3.3 Automatisation grâce à l'intelligence artificielle

Depuis quelques années, les projets et recherches autour de l'utilisation de l'intelligence artificielle (IA) se multiplient, notamment avec la technologie du *Natural Language Processing*. Ainsi The National Archives (UK) ont lancé un grand projet de recherche *Using AI for Digital Records Selection in Government* qui a notamment testé un certain nombre d'outils existants et dont les résultats sont compilés dans différents rapports (The National Archives UK 2016 ; 2020a ; 2021b).

Roland et al. (2019), dans « More human than human ? Artificial intelligence in the archive » donnent une excellente vue d'ensemble des projets utilisant l'IA. L'article de Seth van Hooland et Mathias Cockelbergs, « Unsupervised Machine Learning for Archival Collections : Possibilities and Limits of Topic Modeling and Word Embedding » (2018), présente bien l'utilisation du *NLP* en archivistique, étude de cas à l'appui, tandis que Tim Hutchinson liste de manière très complète les projets et outils conçus pour les archivistes ayant recours au *NLP* (2020).

1.4 Définition des concepts principaux

1.4.1 L'automatisation

Annoncée comme solution incontournable pour les principaux défis de l'archivistique et au centre de notre projet de recherche, le concept même d'*automatisation* est pourtant peu explicité dans la littérature spécialisée – la nuance avec le concept voisin de *systématisation* n'étant d'ailleurs pas toujours claire.

Nous pouvons certes nous accorder sur une définition générale : « Une automatisation est une technique ou un ensemble de techniques ayant pour but de réduire ou de rendre inutile l'intervention d'opérateurs humains dans un processus où cette intervention était coutumière. [...] Elle tend donc à économiser l'intervention humaine sous toutes ses formes » (Encyclopædia Universalis, 2022). La systématisation « consiste à concevoir et à formaliser le processus opérationnel d'une fonction » (Makhlouf Shabou 2015, p. 200). Ce qui signifie qu'une intervention humaine est toujours nécessaire en amont. Pensées en lien avec les différentes fonctions archivistiques, ces définitions ont une utilité limitée.

Les National Archives and Records Administration (NARA) publient en 2014 des directives sur la gestion des documents électroniques : *Automated Electronic Records Management Report/Plan*, un document faisant office de jalon (Hooland, Coeckelbergs 2018). Les NARA y distinguent cinq approches[2] (ou pourrait dire « degrés ») d'automatisation (National Archives and Records Administration 2014, p. 10-14), incluant la systématisation. Certes le document se concentre sur la capture et la classification initiale dans un calendrier de conservation, mais les approches décrites restent néanmoins pertinentes :

- *No automation* – Pas d'automatisation : une gestion « manuelle » des documents électroniques.
- *Rule-based automation* – Automatisation basée sur des règles : une gestion efficace et cohérente des documents peut être réalisée grâce à l'utilisation de règles métier (*business rules*) automatisées qui agissent sur les métadonnées, sur les rôles des utilisateurs ou une autre caractéristique du document.
- *Business Process and Workflow Automation* – Automatisation des processus d'affaires et des *workflows* : au sein de grandes structures d'administration et pour les processus d'affaires bien structurés, les systèmes d'information peuvent être conçus de manière à soutenir les flux d'information tout au long des

[2] « *"approach" as a technical strategy for automating electronic records management* […] » (National Archives and Records Administration 2014, p. 8)

processus et capturer les métadonnées nécessaires pour définir leur sort après leur durée d'utilité administrative et/ou légale.

- *Modular Re-usable Management Tools* – Outils de gestion modulables et réutilisables : une approche intégrée qui fournit des outils, services ou applications de gestion des documents modulables, accessibles et interopérables, offrant ainsi aux services de sélectionner uniquement les outils et modules dont ils ont besoin, économisant ainsi passablement de moyens.
- *Autocategorization* – Classement automatique : l'automatisation la plus avancée où l'analyse informatique du contenu des documents leur attribue une catégorie choisie. Elle passe par exemple par le *machine learning* (également appelée *predictive coding*), les experts entraînent le système à reconnaître les documents pour chaque catégorie choisie (du calendrier de conservation par exemple) en se basant sur des jeux de données. L'entraînement est complété par un processus itératif réalisé sur d'autres documents encodés par la machine elle-même.

Dans notre projet de recherche, nous aborderons donc l'automatisation des fonctions archivistiques comme un « spectre ».

1.4.2 Les fonctions archivistiques

Si leurs conceptions sont multiples, Carol Couture et un collectif d'auteurs ont traité les fonctions archivistiques qui structurent le travail de l'archiviste dans un ouvrage qui fait référence : *Les fonctions de l'archivistique contemporaine* (1999). Nous nous baserons ainsi principalement sur cet ouvrage pour leurs définitions. Si les formats ont évolué, leurs définitions ont peu changé (Kecskeméti, Körmendy 2014 ; Couture, Lajeunesse 2014). Couture dénombre ainsi huit fonctions : l'analyse des besoins ; la création ; l'évaluation ; l'accroissement (l'acquisition) ; la classification ; la description et l'indexation ; la diffusion ; et la préservation. L'analyse des besoins et la création ne seront pas traitées dans ce travail pour des raisons qui seront explicités dans la méthodologie (voir 2.2.4).

Souvent considérée comme le cœur de l'archivistique, l'**évaluation** est :

> *L'« acte de juger des valeurs que présentent les documents d'archives (valeur primaire et valeur secondaire) et de décider des périodes de temps pendant lesquelles ces valeurs s'appliquent auxdits documents dans un contexte qui tient compte du lien essentiel existant entre l'organisme (ou la personne) concerné et les documents d'archives qu'il (elle) génère dans le cadre de ses activités » (Couture 1996, p. 3)*

L'**accroissement** est « […] l'ensemble des mesures employées afin d'accroître le nombre de fonds d'archives d'un organisme pour en permettre l'exploitation » (Lambert, Coté 1992). On distingue deux types d'accroissement : le **versement**, qui est l'opération par laquelle la conservation d'archives passe de l'administration d'origine à un centre d'archives. Le versement traduit la notion de continuum dans la gestion des archives, recouvrant l'ensemble du cycle de vie des documents. D'autre part, l'**acquisition**

regroupe les modes d'accroissement autres que le versement, tels l'achat, le don, le dépôt, le legs, ou l'échange (Couture 1999, p. 147).

La **classification** (composante intellectuelle), qu'on distingue du classement (composante matérielle), est le processus intellectuel d'identification et de regroupement systématique d'articles semblables, d'après des caractéristiques communes pouvant faire par la suite l'objet d'une différenciation si la quantité l'exige. L'objectif ultime de la classification est de faciliter le repérage de l'information que contiennent les documents d'archives. Le principe du fonds doit présider à la constitution des systèmes de classification (Couture 1999, p. 18-20).

La **description** et l'**indexation** sont les opérations qui permettent de décrire les caractéristiques physiques et le contenu des archives. La fonction est régie par trois principes de base : le respect du fonds d'archives ; le fait que l'archiviste doit faire refléter les niveaux de classement (centre d'archives, fonds, série, dossier et pièce) ; et le fait de procéder du général au particulier. Intimement liée à la classification, la description poursuit également l'objectif de rendre accessible l'information contenue dans les documents d'archives ; elle suppose donc une bonne connaissance des besoins des différents utilisateurs. La fonction de la description a peut-être évolué plus que les autres ces dernières années, influencée par les évolutions technologiques – il suffit d'évoquer l'analyse des textes assistée par ordinateur, les systèmes d'indexation ou les documents structurés XML et HTML (Couture 1999, p. 19-20).

La **diffusion** « consiste soit à transmettre à l'utilisateur les informations dont il a besoin, soit à lui donner la possibilité d'y accéder. » (Guinchat, Menou 1981, p. 257). Couture ajoute l'aspect de mise en valeur (1999, p. 22). D'autre part, on considère la diffusion comme « l'objectif ultime » de l'archiviste : « Ce n'est pas une fin en soi d'acquérir, de traiter et de conserver des archives […] elles ne pourront jouer pleinement [leur] rôle que si elles sont adéquatement diffusées. » (Couture, Rousseau 1982, p. 257). Enfin, la mission de la diffusion doit aussi prendre en compte les questions de législation ainsi que les réglementations en matière d'accès et de protections de l'information.

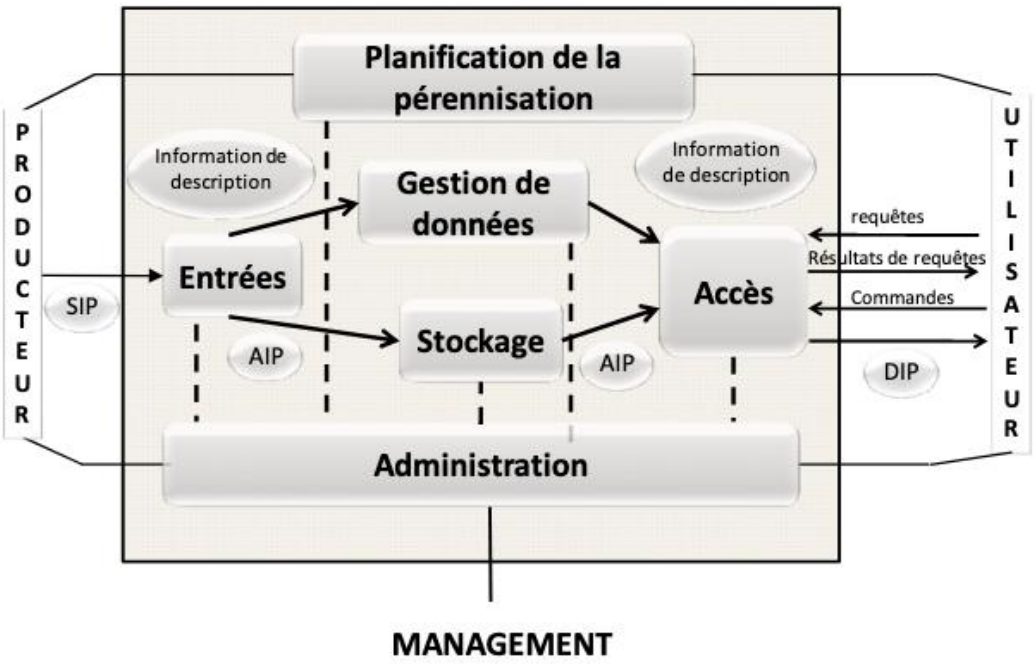
Le groupe de recherche InterPARES définit la **préservation** comme « *The whole of the principles, policies, rules, strategies, and activities aimed at prolonging the existence of an object by maintaining it in a condition suitable for use, either in its original format or in a more persistent format, while leaving intact the object's intellectual form.* » (InterPARES 2018)

1.4.3 Système ouvert d'archivage d'information (OAIS)

Un Système ouvert d'archivage d'information (*Open Archival Information System* – OAIS) est un modèle conceptuel de gestion, de conservation et de préservation à long terme de documents numériques (Makhlouf Shabou 2021), « une organisation […] constituée d'une équipe et de systèmes, dont la responsabilité est de pérenniser des informations et de les rendre accessibles à une Communauté d'utilisateurs cible. » (Comité Consultatif Pour les Systèmes de Données Spatiales (CCSDS) 2017, p. 1.1)

Le modèle OAIS identifie quatre rôles principaux : « archive », « management », « producteurs » et « utilisateurs ». Et cartographie six grands domaines fonctionnels ou entités : « Entrées », « Planifications de la préservation », « Stockage », « Gestion des données », « Administration » et « Accès » (voir Figure 1) chaque entité jouant un rôle[3] spécifique. Ainsi l'entité « Entrée » reçoit, contrôle et valide les objets tandis que l'entité « Stockage » assure la conservation physique des objets ou l'entité « Accès » regroupe les services qui sont en interface directe avec les utilisateurs (Centre Informatique National de l'Enseignement Supérieur (CINES) [s.d.]).

Figure 2 : Entités fonctionnelles OAIS



(Comité Consultatif Pour les Systèmes de Données Spatiales (CCSDS) 2017, p. 4.1)

[3] Pour les rôles détaillés joués par ces entités, voir (Comité Consultatif Pour les Systèmes de Données Spatiales (CCSDS) 2017, p. 4-1 à 4-3)

2. Méthodologie

2.1 Typologie de la recherche

Comme la majorité des études dans le domaine des Sciences de l'information, notre recherche, par son objectif final qui est de proposer une solution pratique à un problème concret – à savoir maîtriser l'augmentation exponentielle de la masse documentaire – est de type *appliqué* (Deschamps 2010). En outre, notre recherche est de type *qualitatif*, puisque les données récoltées ne sont pas de nature statistique ou numérique, mais bien narrative et descriptive. Nous tentons de décrire et d'explorer un phénomène par une étude documentaire et des études de cas, des méthodes de collecte de données essentiellement utilisées dans les recherches qualitatives.

Enfin, notre recherche sera descriptive et exploratoire, la question pivot qui guide notre réflexion étant : quelles sont les possibilités offertes par l'automatisation des fonctions archivistiques pour le traitement des données textuelles ? Bien que l'on tente de cerner le processus de l'automatisation depuis près de soixante ans, aucun compromis ne semble s'être encore dessiné ; le domaine de l'automatisation en archivistique est donc bien un phénomène en cours de maturation qui doit encore être décrit, clarifié et approfondi.

2.2 Étude documentaire

Dans l'optique de disposer d'un panorama général des outils (compris comme des utilitaires, logiciels ou applications web, facilitant l'exécution d'une tâche ou offrant des fonctionnalités (Grand dictionnaire terminologique 2012) actuellement disponibles pour la gestion des archives électroniques et de cerner au plus près les possibilités actuellement offertes par l'automatisation des fonctions archivistiques, nous avons adopté l'approche dite de l'« étude descriptive qualitative » qui consiste à fournir une « description qualitative sommaire des données organisées autour d'un thème ciblé » (Fortin 2016, p. 200). Grâce à une méthode de collecte basée sur le « recueil de textes » (Fortin 2016, p. 202), nous avons réuni deux corpus distincts de documents sur lesquels nous avons mené des analyses successives.

2.2.1 Choix du corpus et collecte des documents

Un corpus documentaire primaire de type « contrasté » (soit « constitué d'énoncés provenant d'auteurs qui ont des options, des préconceptions, des points de vue différents à propos d'une notion ou d'un événement » (Van der Maren 1996, p. 136)) a été constitué via une recension des écrits et une recherche bibliographique thématique : il est composé de publications scientifiques et d'études de cas relatives à

l'automatisation des fonctions archivistiques. Cette « littérature blanche » nous a permis de repérer les outils disponibles et qui ont fait l'objet de l'attention des chercheur·e·s : une liste a ainsi été établie contenant les outils présentés, mentionnés ou utilisés dans le cadre des études précédemment réalisées et des articles publiés jusqu'au 31 décembre 2021. À partir de cette liste d'outils, dont seuls les noms ont été retenus dans un premier temps, nous avons constitué un second corpus documentaire, objet d'une analyse détaillée.

Un corpus documentaire secondaire a été constitué sur la base des lectures réalisées (corpus primaire) et des outils référencés. Ce corpus est composé des pages internet des sites de développeurs et de producteurs des outils listés. Aux pages internet officielles (soit les sites d'entreprises proposant des logiciels ou des solutions informatiques générales pour la gestion de l'information électronique et les sites d'institutions ou de particuliers proposant des outils développés dans le cadre de projets ou de collaborations) se sont ajoutées également, pour les outils en *open source*, les pages internet qui leur sont dédiées sur la plateforme collaborative *GitHub*. Nous avons également eu recours à de la documentation fournie par les développeurs *sur* leur site internet. Cette « littérature grise » comprend, entre autres, les prospectus synthétiques de présentation (type : « *Summary sheets* ») ou encore les manuels d'utilisation – pour autant que ces informations soient accessibles librement (sans qu'il soit donc nécessaire de se créer un compte, de s'inscrire, de s'entretenir avec un·e consultant·e ou de s'acquitter d'un abonnement).

Plusieurs facteurs ont déterminé le choix et le périmètre (ampleur et nature) de ce corpus : garantir un *traitement uniforme* dans la récolte d'informations pour tous les outils référencés ; garantir un *traitement égalitaire* aux fournisseurs d'outils ; garantir *la faisabilité de l'étude* en limitant le nombre de ressources à analyser et à traiter ; se tenir *au plus près de la réalité du terrain*, puisque les praticiens ne s ne disposent le plus souvent que des informations publiques et des textes de présentation des outils avant de sélectionner l'outil qu'ils utiliseront. La collecte des documents (sites internet des développeurs) a été menée entre le 1er août 2021 et le 31 décembre 2021 par les trois auteur·e·s de cette étude.

2.2.2 Collecte des informations à partir d'un corpus documentaire

Pour analyser ce volumineux corpus (environ 70 outils), nous avons conçu une grille d'analyse permettant, pour chaque outil listé, de collecter des informations de nature identique et pré-normalisées :

- **Nom** : intitulé complet de l'outil suivi de l'acronyme utilisé entre parenthèses

- **Accessibilité** : conditions d'accès et d'utilisation de l'outil (libre / propriétaire)
- **Développeur** : nom de l'entreprise ou de l'institution qui a développé l'outil
- **Année de début / année de fin** : date du lancement opérationnel de l'outil et date de la fin de sa maintenance si l'outil n'est plus mis à jour ou s'il n'est plus accessible au téléchargement
- **Fonctionnalités** : liste des « possibilité[s] de traitement offerte[s] par un système informatique, un logiciel ou un site Web » (Grand dictionnaire terminologique 2012)
- **Description** : extrait de la présentation générale de l'outil (sous la forme d'une citation, avec URL exacte et date de consultation) sur le site internet du développeur ou sur la plateforme *GitHub.*
- **Liens** : adresses URL exactes des sites des développeurs ou de la plateforme *GitHub* (avec date de consultation) sur lesquels ont été trouvées les informations récoltées et enregistrées dans la grille de lecture.

2.2.3 Traitement des informations : normalisation des fonctionnalités

Au terme de la récolte des informations extraites du corpus documentaire, nous avions une liste de 235 fonctionnalités distinctes. Les intitulés de ces fonctionnalités provenaient alors *exclusivement* des sites internet des développeurs et pouvaient donc varier considérablement d'un fournisseur à l'autre. Le niveau de détail (granularité) pouvait s'avérer en effet très différent entre les fournisseurs affichant toutes les spécificités techniques de leurs produits et ceux insistant sur la réalisation globale d'une tâche ; le relevé des fonctionnalités comprenait en outre des intitulés en français et en anglais, renvoyant parfois à des traitements identiques (traduction littérale), mais qui pouvaient également refléter des différences plus profondes (en termes conceptuels notamment, entre les cultures archivistiques francophone ou anglophone) ; enfin, le traitement automatique de cette liste via le logiciel *Excel* faisait apparaître comme deux fonctionnalités distinctes des intitulés aux différences linguistiques minimes (via l'utilisation de conjonctions de coordination, de substantifs au lieu de verbes, de formulations au pluriel ou au singulier, etc.).

Pour toutes ces raisons, il a été nécessaire de se livrer à un important travail de *préparation* et de *codage* des données récoltées, en suivant les étapes suivantes :

- **Nettoyage des données :** il s'agissait dans un premier temps d'obtenir les données les plus cohérentes et uniformes possibles en traduisant les termes en langues étrangères, en distinguant les différentes actions dans des assertions contenant plusieurs verbes distincts, en uniformisant les formulations (remplacement des verbes d'action par des substantifs), etc.
- **Réduction des données :** il s'agissait de réunir autant que possible les fonctionnalités qui mettaient en œuvre un traitement informatique identique puisque les doublons ou les intitulés distincts mais très proches en termes de contenu gonflaient artificiellement la liste des fonctionnalités.

- **Traduction des données :** il s'agissait, pour chaque fonctionnalité unique identifiée, de trouver une formulation directement compréhensible par les utilisateurs et utilisatrices futur·e·s de notre grille d'analyse (les archivistes professionnel·le·s) sans trahir les spécificités techniques chères aux informaticien·ne·s.
- **Codage des données :** il s'agissait, après avoir sélectionné une formulation *univoque* pour chaque fonctionnalité isolée, de l'attribuer aux fonctionnalités originalement listées : à chaque intitulé original correspondait donc une fonctionnalité « normalisée ».

Cette opération de *normalisation* permettait alors de disposer d'un langage commun : toutes les fonctionnalités relevées sur les sites internet des développeurs d'outils pouvaient ainsi être mises en commun dans un vocabulaire unifié. Ce travail a permis de réduire le nombre de fonctionnalités de 72 % : des 235 fonctionnalités initiales, l'opération de normalisation a permis d'isoler 66 fonctionnalités uniques. La liste des fonctionnalités normalisées (avec une brève description) est détaillée dans la section 3.2.2.

2.2.4 Conception d'une grille d'analyse pour les fonctions archivistiques

Dans le but de cerner au mieux quelles sont les possibilités offertes par l'automatisation pour la réalisation des fonctions archivistiques, il a été nécessaire de décomposer lesdites fonctions en une suite logique de tâches à accomplir, chaque tâche étant comprise comme un « travail déterminé que le titulaire d'un poste [un·e archiviste dans notre cas] doit exécuter et qui correspond à la division d'une activité spécifique [en l'occurrence : une fonction archivistique] » (Grand dictionnaire terminologique 2012). Cette division d'une fonction archivistique particulière en une suite de tâches que doit réaliser un·e archiviste pour remplir les missions patrimoniales qui lui incombent permettait de rendre *opérationnelles* des définitions trop générales et pas assez concrètes.

En se basant sur le livre de Carol Couture, *Les fonctions de l'archivistique contemporaine* (1999), nous avons pu isoler une série de tâches individuelles. À titre d'exemple, la fonction « Évaluation » a été décomposée en six tâches, chacune formulée grâce à un verbe d'action : « Définir le contexte de création » ; « Acquérir une connaissance de l'institution ou de la personne qui a géré les archives » ; « Élaborer les critères d'évaluation » ; « Attribuer une valeur aux documents » ; « Décider du degré de conservation » ; « Définir des règles de conservation ».

Les fonctions « Analyse des besoins », « Création », « Accroissement » et « Préservation » présentent quelques particularités. Les deux premières n'ont pas été retenues dans notre démarche. « L'analyse des besoins » a été écartée pour des raisons

évidentes car elle apparaît comme « para-documentaire » : l'archiviste analyse les enjeux d'une institution pour proposer diagnostics et remèdes. Pour ce qui est de la fonction « Création », l'archiviste joue « uniquement » un rôle de conseil, le premier responsable de la création de l'information étant bien son créateur (Couture 1999, p. 80). La fonction « Accroissement » a été décomposée en deux fonctions distinctes : « Versement » et « Acquisition », puisque les tâches à accomplir pour l'archiviste peuvent être très différentes s'il s'agit d'archives institutionnelles, dont le versement « traduit la notion de continuum dans la gestion des archives » (Couture 1999, p. 147), ou d'archives non-institutionnelles puisque l'acquisition « implique une cassure dans la gestion des documents » (Couture 1999, p. 147). Quant à la fonction « Préservation », elle a fait l'objet d'un traitement particulier. Comme l'expliquent Couture et Lajeunesse en 2014, « dans *Les fonctions de l'archivistique contemporaine* [1999], le chapitre sur la préservation […] s'intéresse presque exclusivement à la préservation et à la restauration des documents sur supports traditionnels, papier et analogique » (2014, p. 167). Si certaines tâches très générales peuvent être communes à la préservation analogique et numérique et que l'on puisse sans difficulté trouver des équivalents d'un domaine à l'autre (comme pour les tâches suivantes : « Évaluer la situation : lieux, installations, mobilier, documents » ; « Dresser l'inventaire des besoins spécifiques » ; « Contrôler les conditions de stockage et de consultation » ; « Changer de support : copies, fac-similés, microfilms, numérisation » ou encore « Prévenir les désastres et rédiger un plan d'urgence / d'intervention »), il nous a semblé plus pertinent de prendre appui sur le modèle OAIS (défini au point 1.4.3) qui accorde une large place à la préservation numérique et qui est bien diffusé et mis en pratique dans les institutions patrimoniales. Nous nous sommes donc basés sur la traduction française de 2017 du modèle OAIS proposé par le CCSDS et plus particulièrement sur les tâches réalisées par l'entité « Stockage », qui « assure les fonctions et services relatifs au stockage, à la maintenance et à la récupération des AIP » (Comité Consultatif Pour les Systèmes de Données Spatiales (CCSDS) 2017, p. 4.2).

Ce travail de décomposition pratique des fonctions archivistiques a abouti à la formalisation de 37 tâches individuelles. Chaque tâche met en œuvre une action particulière (ou une succession d'actions extrêmement proches) et son champ d'application est défini dans la section 3.2.1.

2.2.5 Analyse et exploitation des résultats

Au terme de la récolte des données (informations sur les outils, collectées sur les sites internet des développeurs), de la normalisation des fonctionnalités informatiques existantes, et de la décomposition des fonctions archivistiques en tâches, nous

disposions de tous les éléments nécessaires pour évaluer les possibilités offertes par l'automatisation pour la réalisation des fonctions archivistiques.

Un tableau à double entrée, avec d'un côté les fonctionnalités (dans l'ordre alphabétique) et de l'autre les tâches (regroupées par fonction, dans l'ordre présenté par Carol Couture dans *Les fonctions de l'archivistique contemporaine*) a été créé afin que nous puissions évaluer quelle(s) fonctionnalité(s) soutenai(en)t la réalisation d'une tâche archivistique. Si la fonctionnalité en question **soutenait** (ou **aidait**) l'archiviste dans l'accomplissement de cette activité spécifique, alors la réponse était *positive* et un chiffre « 1 » était inséré dans le tableau, selon un codage binaire/booléen. Cet exercice d'évaluation a été mené individuellement par les trois membres du projet de recherche et les cas limites ont été discuté *a posteriori* afin de trouver un accord interjuge lors de séances d'échange et de confrontation d'idées et d'expériences.

2.3 Études de cas

Au terme de l'étude documentaire, nous avons procédé à trois études de cas. Cette méthode, qui peut 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 permet de compléter l'étude documentaire par une expérience pratique et une mise en situation réelle. Les fonctionnalités telles que proposées par les développeurs sur leur site internet sont parfois très prometteuses et les formulations utilisées sont tour à tour trop générales, trop précises, ou trop sibyllines. Il s'agit ainsi, grâce aux études de cas, de décrire de manière plus détaillée et de *tester* les fonctionnalités proposées ; de déterminer quel est leur champ d'application *réel* ; d'identifier quelles sont les difficultés auxquelles un·e archiviste est confronté·e au quotidien dans la prise en main des outils disponibles et le traitement d'archives numériques volumineuses ; et enfin de discuter, sur une base empirique, des possibilités d'automatisation des fonctions archivistiques.

2.3.1 Population et échantillon

La population de notre recherche est constituée par les archives numériques produites par des institutions et des organisations privées et publiques qui n'ont pas encore fait l'objet d'un traitement archivistique complet (classification, description, évaluation, préservation).

Ces documents numériques en attente de traitement étant extraordinairement nombreux et volumineux (Coutaz 2016), il est nécessaire de sélectionner, au sein de cette

population première, une population cible constituée par les archives numériques privées produites par des entreprises et des organismes privés, des familles ou des individus. Contrairement aux archives publiques, dont le traitement est régi par différentes lois sur l'archivage[4], imposant entre autres que les documents soient proposés aux institutions patrimoniales avant leur destruction, les archives privées ne sont pas soumises à un cadre légal aussi précis et rien ne contraint les détenteurs privés d'archives de les conserver et de les mettre à disposition (Coutaz 2007). Cette distinction fondamentale a des conséquences profondes qui conditionnent les études de cas réalisées dans le cadre de notre recherche :

- Les archives privées sont conservées par des acteurs hétérogènes (par les organismes eux-mêmes, par des dépôts spécialisés, ou par des institutions patrimoniales publiques) et sont intégrées aux collections via des dons, des dépôts, des achats ou des legs. Il est donc nécessaire de travailler en étroite collaboration avec les producteurs d'archives ou avec leurs ayants droit.
- Les archives privées sont souvent proposées directement pour la conservation définitive et n'ont pas fait l'objet de procédures préalables et systématiques (« analyse des besoins », « capture », etc.) quant à leur création, leur évaluation ou leur organisation. Le relatif *désordre* qui peut en résulter (les données sont souvent moins structurées et moins riches ; les archives privées sont moins bien décrites et classifiées que les archives publiques) rend très certainement l'automatisation d'autant plus nécessaire (et peut-être plus compliquée ou plus partielle).

Au sein de cette population cible (*archives privées*), nous avons sélectionné un échantillon sur lequel nous avons concentré nos réflexions et mené des études de cas. Il s'agit donc d'un échantillon non-probabiliste intentionnel (ou « par choix raisonné », fréquent dans le cadre d'une recherche qualitative).

Les critères de sélection de l'échantillon pour les études de cas sont les suivants :

A. La **faisabilité de l'étude** doit impérativement être prise en compte : outre les aspects juridiques entravant parfois l'accès aux documents (voir ci-dessous, B.), le fonds d'archives doit pouvoir être mis à la disposition des chercheurs durant toute la durée de l'étude (le fonds se devra d'être momentanément clos) par l'institution patrimoniale publique qui a la charge de sa conservation ou directement par le producteur des documents (organismes, entreprises, sociétés).
B. Les **questions légales** soulevées par l'utilisation d'archives privées doivent être clairement formulées et l'accès aux documents doit être autorisé par leur producteur. Sachant qu'aucune base juridique ne contraint les producteurs d'archives privées à conserver leurs documents, à les proposer à une institution

[4] À titre d'exemple la loi fédérale sur l'archivage (LAr ; 152.1) ou les lois cantonales tel la loi sur les archives publiques (LArch ; rsGE B 2 15) du canton de Genève et la loi sur l'archivage (LArch ; 432.11) du canton de Vaud.

publique, ou à permettre aux usagers et chercheurs d'y accéder, nous avons pris garde à sélectionner un fonds d'archives librement accessible (vis-à-vis des données personnelles sensibles notamment) ou choisi en accord avec le producteur ou ses ayants droit.

C. Le **nombre de documents numériques** (données) doit être suffisamment conséquent pour que des tests puissent être menés et que les conclusions tirées soient significatives. On a donc pris garde à choisir un fonds d'archives privées regroupant une quantité de documents numériques suffisante pour qu'il soit justifié de mettre en place des procédures d'automatisation (la prise en main de solutions informatiques, mais aussi les processus d'apprentissage nécessaires à l'utilisation efficace de l'intelligence artificielle sont souvent coûteux et chronophages ; l'investissement est justifié si les coûts engendrés au début permettent une amélioration du traitement dans le futur et une plus grande efficience à long terme dans la gestion de grandes masses de données).

D. La **nature des documents** doit être, dans la mesure du possible, multiple et hétérogène. Réunis autour d'un point commun (il doit s'agir de données *textuelles*), les documents doivent refléter des activités différentes, avoir des natures diverses et être encodés dans plusieurs formats informatiques distincts. Cette hétérogénéité du fonds d'archives sélectionné est importante pour refléter la situation concrète dans laquelle beaucoup d'archivistes et professionnel·le·s de l'information se trouvent au quotidien, lorsqu'ils doivent répondre à des questions d'usager ou traiter un « vrac numérique ».

Parmi les fonds d'archives privées qui répondent à ces critères, nous avons sélectionné les documents de gestion d'une organisation à but non lucratif encore conservés sur des serveurs informatiques internes. Il s'agit d'archives électroniques de l'ONG *Médecins Sans Frontières* – un acteur institutionnel donc. L'accès à l'échantillon a été rendu possible car l'un des auteurs a été actif professionnellement au sein de l'institution et a gardé de bons contacts personnels. Une Charte de confidentialité a été signée par les deux autres auteurs.

Un échantillon de 5 GB a été sélectionné : il s'agit d'une copie des archives numériques d'une mission de *Médecins Sans Frontières* lorsque celle-ci a été fermée[5]. L'intérêt d'un tel échantillon est qu'il couvre toutes les activités de l'institution, les missions fonctionnant tel un « modèle réduit » de l'association dans son ensemble, avec une arborescence grossière par département. De fait, il s'agit cependant plus d'un vrac numérique avec tout de même une grossière structure, contenant *a priori*, des fichiers de formats très divers ce qui va nous permettre potentiellement d'y « appliquer » quasiment toutes les tâches des diverses fonctions. L'échantillon contenant des

[5] Médecins Sans Frontières (MSF) intervient dans des régions où une population se trouve en détresse (catastrophes naturelles ou humaines, situation de belligérance) et lance des « Missions » avec une unité de coordination au sein de la capitale ou une grande ville. Lorsque la situation de détresse est passée, MSF se retire du pays en question et ferme ses « Missions ».

données sensibles, uniquement des listes de noms de fichiers choisis apparaîtront dans ce travail.

2.3.2 Choix des outils testés

Pour réaliser les études de cas annoncées, nous avons sélectionné des outils actuellement disponibles. Le choix a entre autres été fondé sur les critères suivants :

A. **Prix** : le logiciel devait être idéalement gratuit (*Open Access*), ou utilisable en échange d'un prix modeste (on a donc évité donc de sélectionner un logiciel fonctionnant par abonnement, ou exigeant un prix d'entrée trop conséquent).

B. **Accessibilité** : l'outil devait être facilement accessible et largement diffusé/utilisé, afin de permettre sa prise en main rapide, de favoriser les échanges entre utilisateurs et d'encourager les comparaisons et les retours d'expérience.

C. **Régime juridique** : la propriété des données devait rester en mains publiques (si les données sont la propriété d'une institution patrimoniale publique) ou entre celles du producteur ou de l'ayant droit légal (dans le cas d'une organisation/entreprise privée). Les données ne devaient donc pas être *captées* par le fournisseur (à l'instar de ce qui se passe avec certains réseaux sociaux) et le régime juridique dans lequel se trouvent les serveurs informatiques du fournisseur doit être conforme aux exigences suisses en termes de protection des données[6].

D. **Compétences et services** : l'outil informatique sélectionné devait proposer des formes d'automatisation (selon la définition exposée au point 1.4.1) pour les fonctions archivistiques (décomposées en tâches), dans le cadre du traitement d'archives électroniques privées.

Sur la base de ces critères, et grâce au relevé des outils actuellement disponibles effectué dans le cadre de l'étude documentaire, il apparaissait que les outils suivants étaient particulièrement indiqués : Archifiltre, DROID (Digital Record Object Identifier) et Karen's Directory Printer.

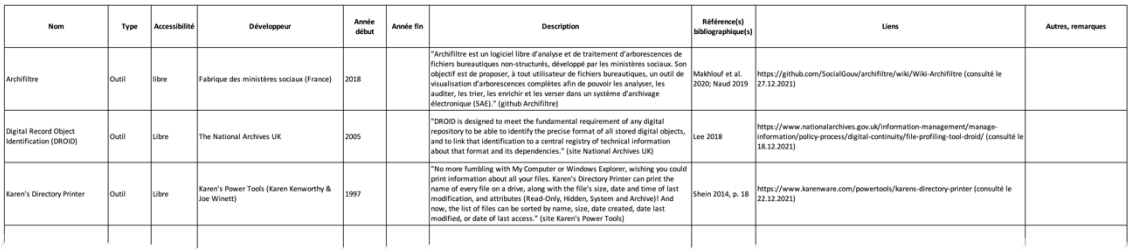
[6] Loi fédérale du 19 juin 1992 sur la protection des données (LPD ; 235.1) et Règlement général sur la protection des données (RGPD ; L 119/1).

3. Résultats

3.1 Tableau synoptique

Les résultats de notre collecte d'information (voir 2.2.2) ont été compilés dans un tableau synoptique (Annexe 1).

Figure 3 : Extrait du tableau synoptique.



Nom	Type	Accessibilité	Développeur	Année début	Année fin	Description	Référence(s) bibliographique(s)	Liens	Autres, remarques
Archifiltre	Outil	libre	Fabrique des ministères sociaux (France)	2018		"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)." (github Archifiltre)	Makhlouf et al. 2020; Naud 2019	https://github.com/SocialGouv/archifiltre/wiki/Wiki-Archifiltre (consulté le 27.12.2021)	
Digital Record Object Identification (DROID)	Outil	Libre	The National Archives UK	2005		"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 National Archives UK)	Lee 2018	https://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/ (consulté le 18.12.2021)	
Karen's Directory Printer	Outil	Libre	Karen's Power Tools (Karen Kenworthy & Joe Winett)	1997		"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 Karen's Power Tools)	Shein 2014, p. 18	https://www.karenware.com/powertools/karens-directory-printer (consulté le 22.12.2021)	

Septante d'entre eux sont effectivement des outils, au sens où nous l'avons défini plus haut, les autres étant des *vocabulaires*, *modèles* ou *guides*, qui, sans accomplir une action, sont nécessaires à l'activité des archivistes. Mais leurs descriptions étaient parfois si générales qu'il était difficile de leur attribuer des fonctionnalités, raison pour laquelle nous avons préféré les regrouper dans un tableau différent (Annexe 2), dans lequel nous avons aussi inclus des *projets*. Certains étant liés à un ou plusieurs outils, il nous a semblé intéressant d'en conserver tout de même une trace. En outre, ces projets nous semblaient être des ressources utiles pour toute personne cherchant à lancer son propre projet d'archivage numérique.

Concernant leur accessibilité, près des deux tiers des outils (47) sont libres. Ce qui est peu surprenant, considérant qu'ils ont été, en grande partie, développés par des institutions publiques (archives nationales ou universités) ou grâce à des financements publics (l'Union européenne, notamment). On en trouve une petite portion qui a été créée par des individus ou groupes d'individus, souvent des professionnel·le·s des Sciences de l'Information ou de l'Informatique.

3.2 Visualisation

Par le traitement des informations, nous avons isolé 37 tâches et 66 fonctionnalités, dont les définitions sont listées plus bas (voir 3.2.1 et 3.2.2).

Dans un second temps, nous avons procédé à la mise en lien de ces tâches et de ces fonctionnalités, en tentant de répondre à la question : cette fonctionnalité soutient-elle cette tâche ?

Pour une réponse positive (selon un mode binaire 1/0, comme expliqué plus haut), il n'était pas nécessaire que la fonctionnalité accomplisse la tâche dans son intégralité mais permette à l'archiviste de la réaliser plus aisément.

3.2.1 Définitions des tâches

Les définitions des tâches sont prioritairement inspirées de l'ouvrage dirigé par Carol Couture (1999).

Évaluation

1. **Définir le contexte de création** : Acquérir une connaissance de l'institution ou de la personne qui a géré les archives. Équivaut à la macro-évaluation où l'on s'intéresse aux raisons pour lesquelles le document existe : aux mandats, fonctions, activités du créateur et où l'on décide quelles unités administratives on va étudier. (Cook 1992 ; Couture 1999)

2. **Élaborer les critères d'évaluation** : Le critère d'évaluation est une caractéristique, un point de repère qui permet de juger de la valeur primaire ou secondaire (Schellenberg 1964 ; 1965) du document. Il n'existe pas de grille applicable à tous les milieux, et il existe de multiples approches différentes « la densité de la documentation n'a d'égal que l'enchevêtrement des notions qu'on y trouve » (Couture 1999, p. 116). On citera à titre d'exemple Boles et Young qui articulent leurs critères en trois modules : valeur de l'information (caractéristiques liées aux fonctions de l'institution par exemple) ; coûts liés à la conservation (achat, transfert, quantité de travail, etc) ; conséquences de la décision résultant de l'évaluation (caractéristiques liées aux relations externes ou aux politiques et pratiques internes de l'institution) (Boles, Young 1991).

3. **Attribuer une valeur aux documents** : On attribue au document soit une valeur *primaire* : administrative, légale ou financière, soit une valeur *secondaire*, qui correspond aux autres utilisations possibles des archives. Schellenberg (1964, 1965) y voit deux composantes : la valeur de *témoignage* lié soit à la structure, à la fonction et aux activités propres à chaque unité d'un organisme donné, soit à l'histoire de l'institution. L'autre composante est le concept plus général d'*information* qui recouvre une dimension extra-institutionnelle (Couture 1999, p. 113).

4. **Décider du degré de conservation** : Il s'agit de trouver un équilibre entre conservation et élimination, en d'autres mots de décider soit d'une conservation intégrale ; soit d'une conservation partielle / d'une élimination partielle ; soit d'une élimination intégrale (Guyot-Jeanning 1984).

5. **Définir des règles de conservation** : La règle de conservation est une norme fixée à partir du jugement que l'on porte sur les valeurs primaires et secondaires que présentent les archives. « En portant ce jugement, l'archiviste fixe la durée de conservation, le cheminement et le traitement des archives depuis leur création jusqu'à leur élimination ou leur versement aux archives définitives » (Couture 1999, p. 117).

Versement (Accroissement)

6. **Inventorier l'ensemble des documents** : Réalisation d'un « inventaire de l'ensemble des documents produits par l'organisme » permettant d' « identifier

les séries de dossiers générés par l'organisme et de les analyser dans leurs interrelations afin de déterminer non seulement lesquelles sont les plus importantes à conserver – action qui relève de l'évaluation –, mais également de fixer les ressources financières, humaines et matérielles requises pour gérer efficacement les transferts au centre de gestion des archives intermédiaires ou les versements au service d'archives. » (Couture 1999, p. 161-162)

7. **Éliminer les archives intermédiaires non conservées** : « L'élimination est entourée d'une série de procédures qui ont pour objectifs d'assurer l'efficacité administrative et de protéger la crédibilité du système de gestion documentaire et l'authenticité des documents versés au service d'archives. […] Avant de procéder à l'élimination, le centre de gestion des archives intermédiaires avise l'unité qui a effectué le transfert que les dossiers mis en liste seront détruits à une date indiquée à moins qu'elle ne s'y oppose. […] [L'] avis de destruction signé, sur lequel est inscrit la date réelle d'élimination, est conservé en permanence afin d'attester que l'action a eu lieu dans le cadre normal des activités de l'organisme. » (Couture 1999, p. 167) « Procédure réglementaire qui consiste à détruire des documents dont la conservation n'est plus justifiée. » (PIAF, 2015)

8. **Verser les archives définitives au service d'archives** : Acte par lequel les documents, « en vertu du calendrier de conservation », « sont versés de plein droit au seul service d'archives autorité à les recevoir. Ce versement implique que l'unité créatrice ne peut plus rappeler les documents, n'est plus responsable de leur conservation physique, n'établit plus les règles de communicabilité et ne gère plus la protection des renseignements personnels qu'ils renferment ; toutes ces responsabilités sont dorénavant assumées par le service d'archives ». Les versements doivent être effectués « dans le cadre normal des activités de l'organisme », « afin de protéger l'authenticité des documents ». Enfin, « l'acte de réception du versement est aussi normalisé à l'aide d'un bordereau d'enregistrement qui témoigne de l'acceptation officielle du versement par l'archiviste, avec les obligations qui en découlent. » (Couture 1999, p. 167-168)

Acquisition (Accroissement)

9. **Établir la liste des fonds souhaités** : « Recherche qui identifie non seulement les fonds de créateurs exceptionnels, mais aussi ceux de personnes représentatives de leur milieu » et « recherche des propriétaires de ces fonds qui peuvent ou non en être les créateurs » ; le fonds est alors fiché et les informations suivantes sont répertoriées : intérêt, contenu possible, références retrouvées, localisations potentielles, contacts établis (Couture 1999, p. 181).

10. **Établir la relation personnelle avec les donateurs** : « Une fois qu'un fonds visé a été localisé, l'archiviste rend visite à son propriétaire afin d'évaluer l'intérêt du fonds. Ce contact personnel constitue la meilleure base de négociation en vue d'une éventuelle acquisition » (Couture 1999, p. 181).

11. **Négocier les acquisitions et signer les ententes** : Discuter avec le propriétaire ou le cédant des conditions d'acquisition des archives : l'archiviste doit évaluer « l'importance qu'accorde le propriétaire à ses documents », « déceler le moment opportun pour faire l'acquisition », connaître le marché et la législation en vigueur (en cas de vente ou de don contre un dégrèvement fiscal par exemple) et informer le cédant des facteurs qui influenceront sa décision (restrictions de consultation, droits d'auteur, etc.). Au terme de la négociation, l'acquisition est « formalisée dans une entente qui a force légale ». (Couture 1999, p. 181-182)

12. **Transférer les documents au service d'archives** : « organiser le transfert physique des documents au service d'archives de manière à maintenir l'ordre original et à assurer un transport sécuritaire » (Couture 1999, p. 181).

13. **Inventorier sommairement les documents** : opérer le « contrôle intellectuel » et élaborer « l'inventaire du fond [sous la forme d'un récolement] à l'aide d'un formulaire normalisé, s'il ne l'a pas été lors de la mise en boîte. » (Couture 1999, p. 181-182)

14. **Annoncer les acquisitions** : informer le public ainsi que les « chercheurs des disciplines les mieux représentées dans le fonds » de l'acquisition des archives via « l'état général du service d'archives, son site Web ou son bulletin » ainsi que dans des publications spécialisées (Couture 1999, p. 182, 214).

Classification

15. **Importer le plan de classification des archives courantes** : Dans le cas de versements, il est recommandé de maintenir le plan de classification des archives courantes avec des ajustements si nécessaires. « En important le plan de classification des archives courantes, l'archiviste respecte le travail de ses prédécesseurs et rend accessibles des informations qui, autrement, demeuraient inutilisées en raison d'un traitement sans cesse différé. » (Couture 1999, p. 248).

16. **Délimiter le fonds d'archives** : Pour délimiter les fonds, on applique le principe du respect des fonds qui fait consensus chez les archivistes, la délimitation se fait en attribuant les documents au créateur·trice des archives. Les critères d'identification des producteurs d'archives sont : « existence juridique du créateur, mandat officiel, position hiérarchique définie, autonomie de fonctionnement et structure fixée dans un organigramme. » (Couture 1999, p. 229).

17. **Situer le fonds dans un plan général de classification** : Chaque institution d'archives élabore un plan général de classification qui englobe tous les fonds d'archives qu'elle détient. Le plan général découle d'une volonté des responsables de regrouper les fonds par ensembles cohérents et logiques dans leur contexte administratif et culturel. (Couture 1999, p. 239-240)

18. **Répartir les documents en fonction de ses créateurs** : L'archiviste répartit les documents en fonction de son ou ses créateur·trices respectif·ves, « […] pour ce faire, il analyse les documents constitutifs et les rapports d'activités de chaque créateur de fonds, dans le cas des organismes, ou les notices biographiques, dans le cas des personnes, afin de déterminer leurs fonctions et leurs activités. » (Couture 1999, p. 223-224)

19. **Choisir un modèle de classification** : Il y a un large consensus pour préserver l'« organicité » et l'unité du fonds, mais on trouve néanmoins des modèles de classifications par régions géographiques, périodes chronologiques ou structures administratives par exemple (Couture 1999, p. 241).

Description

20. **Identifier les usagers et leurs besoins** : « Identifier les catégories d'utilisateurs recourant aux archives et […] cibler leurs besoins d'information tout comme leurs stratégies de recherche ». (Couture 1999, p. 257)

21. **Présenter des caractéristiques physiques des documents** : Présenter les « données objectives que sont les caractéristiques physiques des documents

(date(s) de création, étendue linéaire, état physique, restrictions à la consultation, présence de caractéristiques physiques particulières […]) ». (Couture 1999, p. 265)

22. **Analyser le contenu des documents** : « Présenter sous une forme concise et précise les données caractérisant l'information contenue dans un document ou un ensemble de documents (de la pièce au fonds), du bordereau à l'inventaire. Par extension, le résultat de cette opération. » (Sibille-De Gimoüard et Caya, 2009)

23. **Présenter le contexte de création et d'utilisation** : « Identifier l'unité productrice des documents […] pour permettre d'identifier le créateur du fonds et distinguer les fonds entre eux ». Rédiger une « Histoire administrative ou une notice biographique pour permettre à des utilisateurs de tout horizon de situer les documents décrits dans leur contexte ». (Couture 1999, p. 269)

24. **Concevoir l'instrument de recherche** : Mettre en place un « instrument de description contenant des informations permettant d'établir un contrôle sur les documents et de faciliter leur repérage ». (Couture 1999, p. 276)

25. **Indexer** : Sélectionner « [d]es termes et [d]es expressions d'indexation dans les documents originaux ou dans des résumés » et les transcrire « dans un langage documentaire donné ». (Couture 1999, p. 318)

Diffusion

26. **Former les usagers** : « Rendre la clientèle la plus autonome possible. » (Couture 1999, p. 383)

27. **Développer la clientèle** : « Établir une politique de relations publiques » pour « ven[dre] » [le] système de gestion des archives aux décideurs » et « permettre le maintien de l'intérêt de la clientèle (créateur ou utilisateur) pour la gestion des archives » et « pour justifier les ressources qui leur sont attribuées et prétendre au développement des services qu'ils offrent ». « Augmenter le bassin d'utilisateurs. » (Couture 1999, p. 383-385)

28. **Mettre en valeur les documents** : « Diriger les usagers vers des sujets méconnus susceptibles de les intéresser et des sources d'informations ignorées » (Couture 1999, p. 385).

29. **Mettre à disposition les documents** : « Instaurer des mesures de contrôle des unités de rangement dans les magasins pour assurer le repérage et la sécurité des documents réquisitionnés par les chercheurs ». « Empêcher l'accès, par les chercheurs, à l'information qui permet le repérage des documents frappés de restrictions. » (Couture 1999, p. 403)

30. **Reproduire les documents** : Gérer les commandes de reproduction, les documenter et permettre un contrôle des droits d'auteur (Couture 1999, p. 410).

Préservation numérique (OAIS)

31. **Convertir les formats de fichiers (entité « Entrée »)** : « La fonction <générer un AIP> transforme un ou plusieurs *SIP* en un ou plusieurs *AIP* conformes aux *normes de documentation et de formatage des données* de l'Archive. Elle peut impliquer des conversions de format de fichier, la collecte de l'Information de représentation appropriée, des conversions de représentation des données ou une réorganisation de l'Information de contenu des SIP » (CCSDS 2017, p. 4.7).

32. **Gérer la hiérarchie du stockage (entité « Stockage »)** : fournir « des statistiques d'exploitation qui font l'inventaire des supports à disposition, de la capacité de stockage disponible dans les différentes couches de la hiérarchie de stockage, et des statistiques d'utilisation » (CCSDS 2017, p. 4.10).

33. **Remplacer les supports (entité « Stockage »)** : offrir « la possibilité de reproduire des AIS dans le temps » ; « La stratégie de migration doit consister à choisir un support de stockage en tenant compte des taux d'erreurs réels et attendus caractérisant les différents types de supports, de leurs performances, et de leur coût d'acquisition » ; les types de migration sont les suivants : « Rafraîchissement de support », « Duplication », « Ré-empaquetage » et « Transformation » (2017, p. 4.10, 5.5-5.6).

34. **Contrôler les erreurs (entité « Stockage »)** : garantir « avec une probabilité statistiquement acceptable, qu'aucun composant de l'AIP n'a été corrompu lors d'un quelconque transfert interne des données de l'Entité <Stockage> » (CCSDS 2017, p. 4.10).

35. **Fournir un plan de reprise (entité « Stockage »)** : « mécanisme pour dupliquer les contenus numériques de la collection d'archives et, par exemple, pour stocker la copie dans une installation physiquement séparée. » (CCSDS 2017, p. 4.10)

36. **Gérer la configuration du système (entité « Stockage** ») : assurer « la maîtrise technique du système d'archivage pour surveiller en permanence son fonctionnement global et contrôler systématiquement les modifications de sa configuration » (CCSDS 2017, p. 4.13-4.14).

37. **Tenir une veille technologique (entité « Administration »)** : « suivre les technologies numériques émergentes, les normes d'information et les plates-formes informatiques (c.-à-d. matérielles et logicielles) pour détecter les technologies qui pourraient conduire à une obsolescence de l'environnement informatique de l'Archive et empêcher l'accès à certains fonds courants de l'Archive » (CCSDS 2017, p. 4.17).

3.2.2 Définitions des fonctionnalités

Pour la rédaction de ces définitions, les dictionnaires et glossaires suivants ont été nos inspirations principales :

- Glossaire du Portail International Archivistique Francophone (Portail International Archivistique Francophone 2015)
- Grand dictionnaire terminologique (GDT) de l'Office québécois de la langue française (Grand dictionnaire terminologique 2012)
- Dictionary of Archives Terminology de la Society of American Archivists (SAA Dictionary 2005-2022)

*

1. **Ajout de métadonnées descriptives intellectuelles** : Génération de métadonnées nouvelles au sujet du contenu et du contexte du document selon des standards archivistiques particuliers.

2. **Ajout de métadonnées descriptives techniques** : Génération de métadonnées nouvelles au sujet des caractéristiques techniques du document selon des standards archivistiques particuliers.

3. **Analyse de volumétrie** : Mesure du « poids » des documents numériques et des espaces de stockage disponibles.
4. **Analyse des sentiments (NLP)** : Identification de sentiments dégagés par un texte.
5. **Annotation de document** : Ajout d'une remarque, d'un mot-clé ou d'un *tag* en marge d'un texte par exemple pour une utilisation ultérieure.
6. **Application des délais de conservation** : Élimination ou conservation des documents selon des critères établis par un calendrier de conservation.
7. **Attribution d'identifiants pérennes** : Attribution d'un identifiant généré selon des standards qui « le prémunissent contre toute modification et toute réplication [et] restera toujours identique et unique au monde » (FranceArchives 2021).
8. **Bloquage d'écriture :** Restriction de modification d'un document.
9. **Calcul de la valeur archivistique** : Estimation d'une valeur archivistique par l'attribution et la pondération de mesures quantitatives.
10. **Changement d'extension** : Modification de l'extension pour imposer l'ouverture du document dans un outil spécifique.
11. **Classification et catégorisation de documents (NLP)** : Classement des documents par classes selon une méthode algorithmique.
12. **Comparaison de fichiers texte** : Comparaison du contenu des documents en isolant leurs différences et leurs contenus communs.
13. **Compilation de métadonnées** : Collection de métadonnées issues de sources multiples et réorganisées selon un standard archivistique spécifique.
14. **Contrôle d'autorité** : Identification des personnes et d'autres concepts sans ambiguïté.
15. **Contrôle de conformité du SIP** : Vérification de la conformité aux conditions d'importation des documents.
16. **Contrôle de signature électronique** : Vérification de la validité d'une signature électronique.
17. **Contrôle des droits d'accès** : Application de restrictions d'accès (par exemple, accès utilisateur, délais de protection).
18. **Contrôle du vocabulaire** : Utilisation d'une liste de termes prédéfinis (thésarus, taxonomie, etc.)
19. **Conversion de format** : Transformation d'un format de représentation des données dans un autre.
20. **Cassage de mot de passe** : Supprimer la protection d'un mot de passe.
21. **Création d'arborescence virtuelle** : Création d'un plan de classement qui ne modifie pas l'emplacement des documents sur son support.
22. **Création d'un bordereau d'élimination** : Création d'un document contenant le relevé détaillé des documents éliminés par le service d'archive.

23. **Création d'un bordereau de versement** : Création d'un document contenant le relevé détaillé des documents reçus par le service d'archive.
24. **Création d'un calendrier de conservation** : Création d'un guide regroupant les règles de conservation et d'élimination selon les typologies de documents.
25. **Création d'un plan de classement** : Création d'un système d'organisation des documents.
26. **Création d'un rapport de recherche** : Compilation d'une liste de documents sélectionnés par le biais d'une recherche.
27. **Création d'un rapport de récolement** : Compilation d'une liste d'informations sur l'ensemble des documents d'un lot/fond.
28. **Création d'un rapport synthétique** : Compilation d'une sélection d'informations sur un lot de documents.
29. **Création de copies de téléchargement** : Création d'une copie du document exploitable selon des conditions préétablies (restriction de partage, de modification ; ajout d'information de type empreinte digitale numérique).
30. **Création de copies d'accès** : Création d'une copie du document dans le seul but de consultation (application de restrictions d'accès et de modification).
31. **Décryptage de fichiers** : Décodage d'un cryptogramme (« données textuelles rendues inintelligibles à l'aide d'un algorithme de chiffrement » (Grand dictionnaire terminologique 2012).
32. **Dédoublonnage** : Repérage et élimination de documents exactement similaires.
33. **Évaluation des risques archivistiques** : Analyse des risques auxquels les documents peuvent être exposés.
34. **Exportation vers un SAE** : Déplacement de documents vers un Système d'archivage électronique.
35. **Extraction d'informations contextuelles (NLP)** : Création de métadonnées à partir de l'analyse du contenu d'un document par *Natural Language Processing*.
36. **Extraction de métadonnées** : Sélection, isolation et copie de métadonnées d'un document.
37. **Extraction de texte :** Conversion du contenu textuel d'un document dans un format standard.
38. **Garantie de la valeur probante (empreinte digitale numérique)** : Vérification de l'authenticité d'un document, garantissant son utilité comme preuve légale ou historique (généralement par le biais d'une empreinte digitale numérique).
39. **Gestion d'ontologies** : Création et modification d'ontologies (« ensemble d'informations dans lequel sont définis les concepts utilisés dans un langage donné et qui décrit les relations logiques » (Grand dictionnaire terminologique 2012).
40. **Moissonnage du web** : Extraction de données du web à l'aide d'un robot d'indexation.
41. **Identification de format** : Reconnaissance de type de représentation des données d'un document.

42. **Identification de langue** : Reconnaissance de la langue dans laquelle un contenu a été rédigé.
43. **Importation d'un plan de classement préexistant** : Extraction et copie d'un système d'organisation préexistant au moment de l'importation des documents.
44. **Importation de documents** : Transfert des documents (contenus et métadonnées) du créateur au service d'archives.
45. **Journalisation** : Enregistrement chronologique des différentes opérations subies par un document.
46. **Migration de support** : Transfert de documents d'un support à un autre.
47. **Modèle de description standardisé** : Mise à disposition d'un modèle (*template*) proposant des champs standardisés pour la description archivistique.
48. **Prévisualisation de fichiers** : Affichage de documents sans ouverture de fichier.
49. **Recherche à facettes** : Méthode de recherche basée sur la classification à facettes.
50. **Recherche générale** : Méthode de recherche basée sur plusieurs méthodes de recherches.
51. **Recherche géospatiale** : Méthode de recherche par lieu géographique.
52. **Recherche plein texte** : Méthode de recherche par analyse du contenu textuel des documents.
53. **Reconnaissance d'entités nommées (NLP)** : Extraction d'entités nommées (noms propres, lieux géographiques, dates, entreprises, institutions, etc.) d'un corpus documentaire par *Natural Language Processing*.
54. **Reconnaissance d'images** : Identification du contenu d'une image.
55. **Reconnaissance optique de caractères imprimés** : Identification du contenu textuel d'un document imprimé en vue d'une extraction de texte.
56. **Reconnaissance optique de caractères manuscrits** : Identification du contenu textuel d'un document manuscrit en vue d'une extraction de texte.
57. **Récupération de données** : Rétablissement des données à partir d'un format ou support défectueux ou obsolète.
58. **Remaniement de l'arborescence** : Modification du plan de classement.
59. **Renommage de fichiers** : Modification du nom d'un document selon des standards préétablis.
60. **Sensitivity Review (NLP)** : Identification d'informations sensibles (dont *Personally identifiable information*) par *Natural Language Processing*.
61. **Topic modeling (NLP)** : Génération de *Topic model* (modèles thématiques) sur la base de modèles probabilistes, permettant de déterminer les thèmes abstraits d'un document par clusters (Blei 2012 ; Brett 2012).

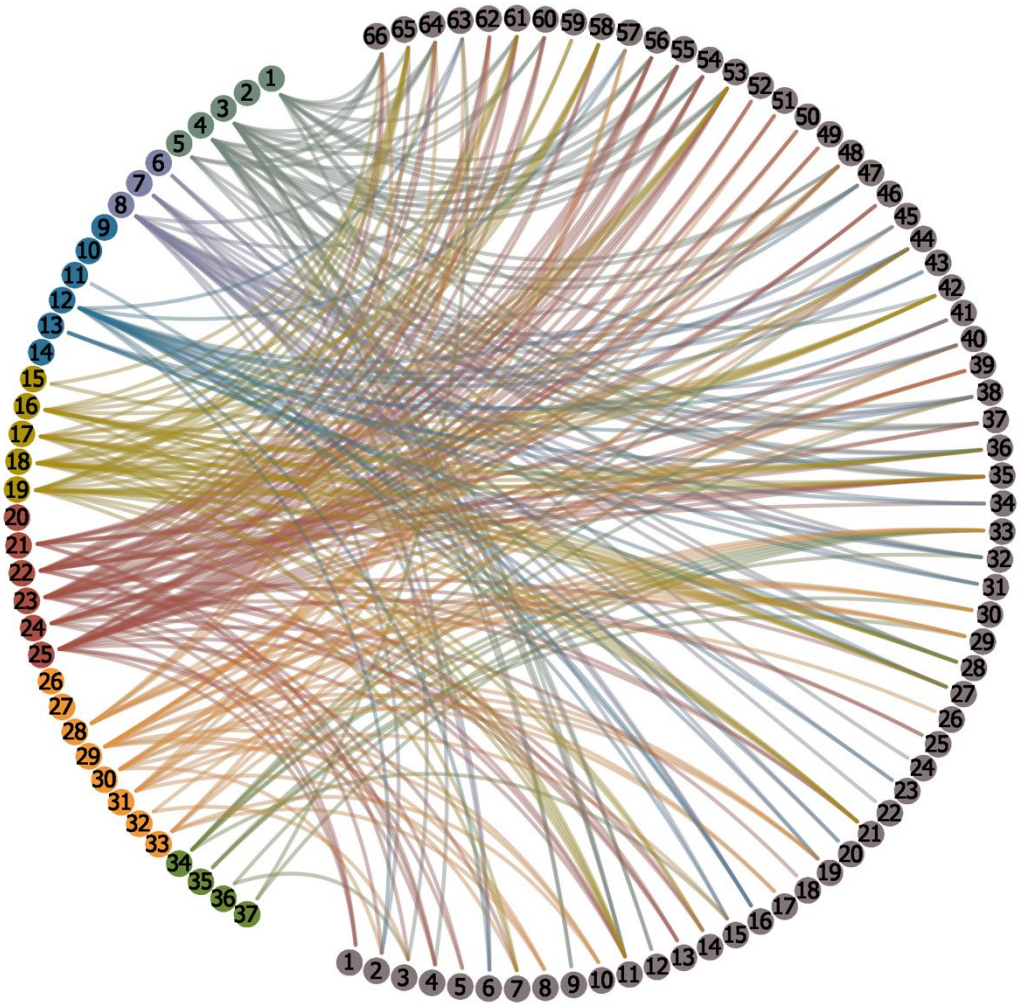
62. **Utilisation de Linked Data** : Liaison entre informations et entités nommées comprises dans le document aux *Linked Open Data* (données ouvertes liées), basée sur les principes Resource Description Format (RDF) du web sémantique.
63. **Validation de format** : Vérification de l'obsolescence des formats.
64. **Visualisation de chronologie** : Représentation graphique des informations chronologiques concernant les documents.
65. **Visualisation de l'arborescence** : Représentation graphique du plan de classement.
66. **Visualisation par cartes géographiques** : Représentation graphique des informations géographiques concernant les documents.

3.2.3 Un outil pour le développeur

Afin de mieux analyser ces données, nous avons créé une visualisation, sous la forme d'un diagramme de membrures (ou *chord diagram*), type de diagramme généralement utilisé pour représenter les relations ou les flux entre les données. Pour ce faire, nous avons utilisé les logiciels de la suite Tableau.

Sur la gauche, 37 *nœuds*, dont la couleur correspond à la fonction archivistique, représentent les tâches, sur le côté droit, les 66 fonctionnalités ont été listées par ordre alphabétique. La numérotation correspond à celle des tâches et fonctionnalités listées plus haut. Chaque arc reliant une fonctionnalité et une tâche correspond à un « 1 » dans notre tableau.

Figure 4 : Représentation graphique des liens entre *tâches* et *fonctionnalités*



Cette visualisation nous permet de mettre en évidence les tâches archivistiques les plus soutenues et celles qui le sont peu, voire pas du tout. On compte ainsi sept tâches non-soutenues sur les 37 : « Élaborer les critères d'évaluation » (n° 2) ; « Établir la liste des fonds souhaités » (n° 9) ; « Établir la relation personnelle avec les donateurs » (n° 10) ; « Annoncer les acquisitions » (n° 14) ; « Identifier les usagers et leurs besoins » (n° 20) ; « Former les usagers » (n° 26) ; « Développer la clientèle » (n° 27). À l'opposé, la tâche « Indexer » (n° 25) est la plus soutenue.

En revanche, ayant choisi d'exclure les fonctionnalités ne soutenant aucune tâche archivistique, dans la volonté d'en réduire un nombre déjà important, toutes soutiennent au moins une tâche. Mais on peut observer que ce sont celles qui exploitent la technologie du *Natural Language Processing* qui tirent leur épingle du jeu.

Figure 5 : La tâche *Indexer*

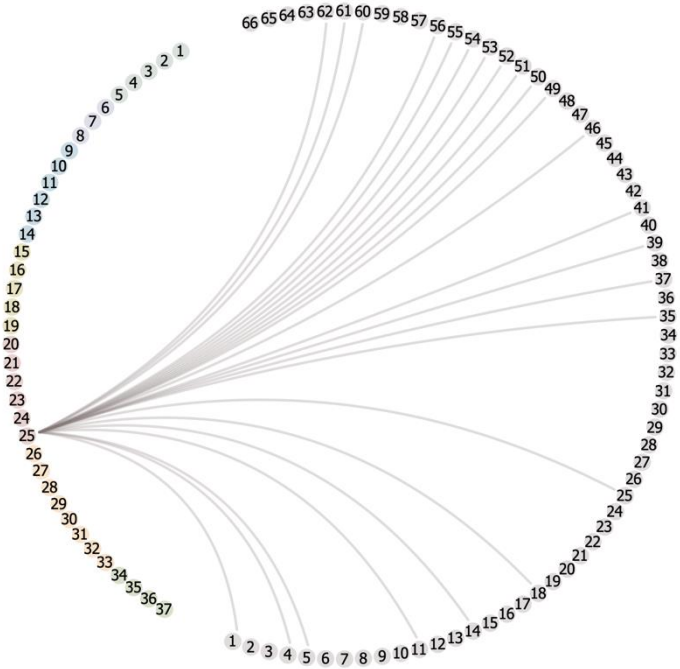
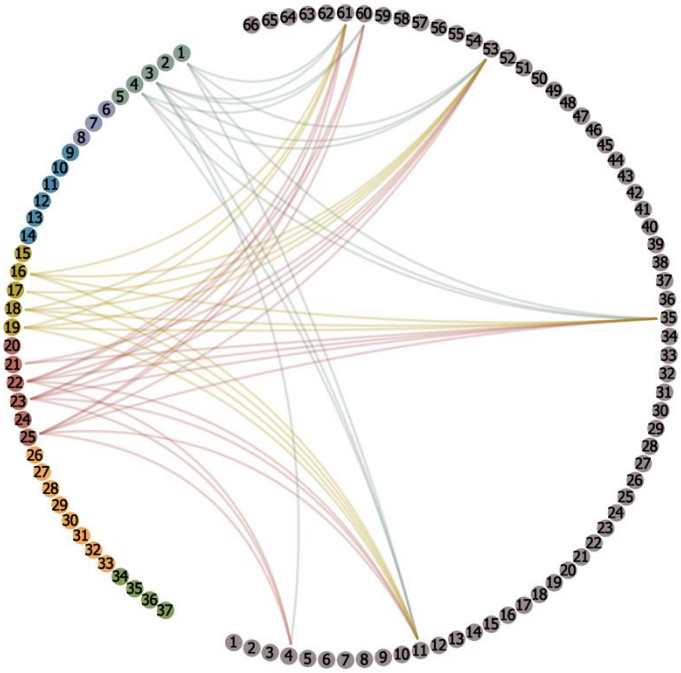


Figure 6 : Les fonctionnalités utilisant le « NLP »



3.2.4 Un outil pour l'utilisateur

Par la suite, grâce à un second tableau, basé sur notre tableau synoptique initial et les fonctionnalités listées pour chaque outil, les informations disponibles peuvent être filtrées, rendant visibles les fonctionnalités de chaque outil.

Figure 7 : Archifiltre

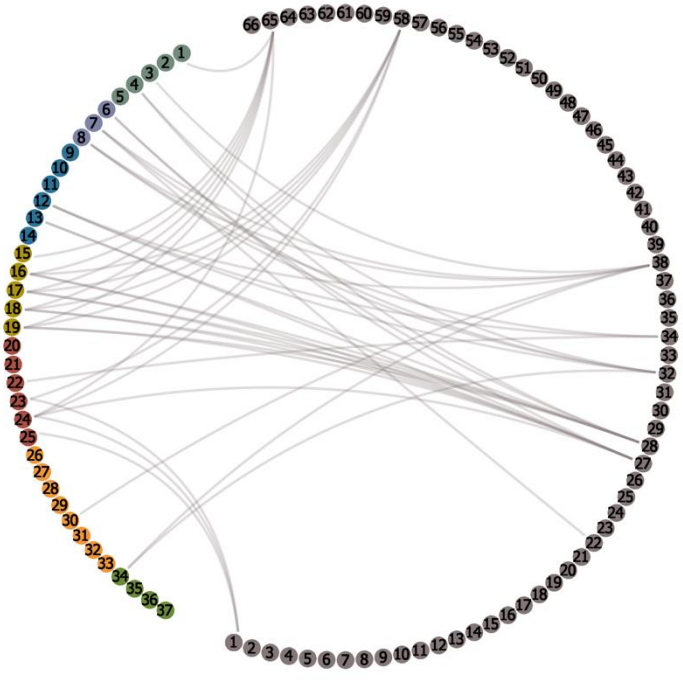


Figure 8 : DROID

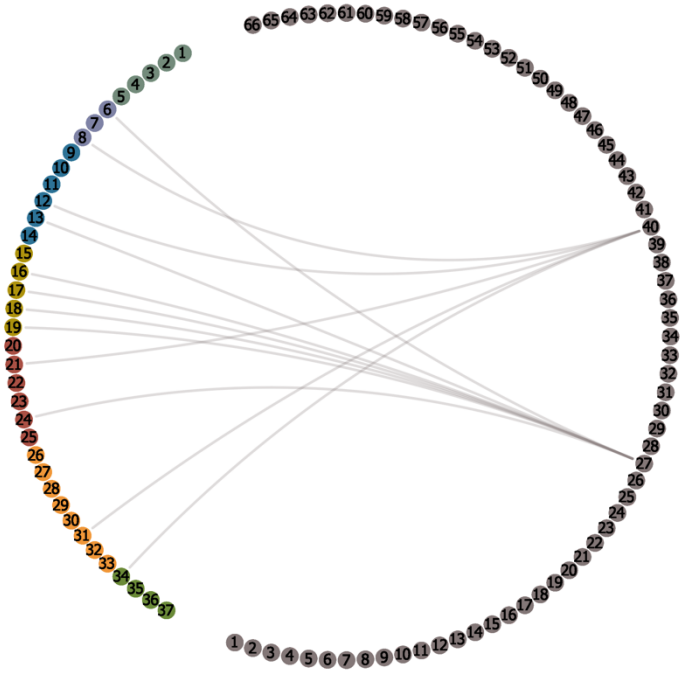
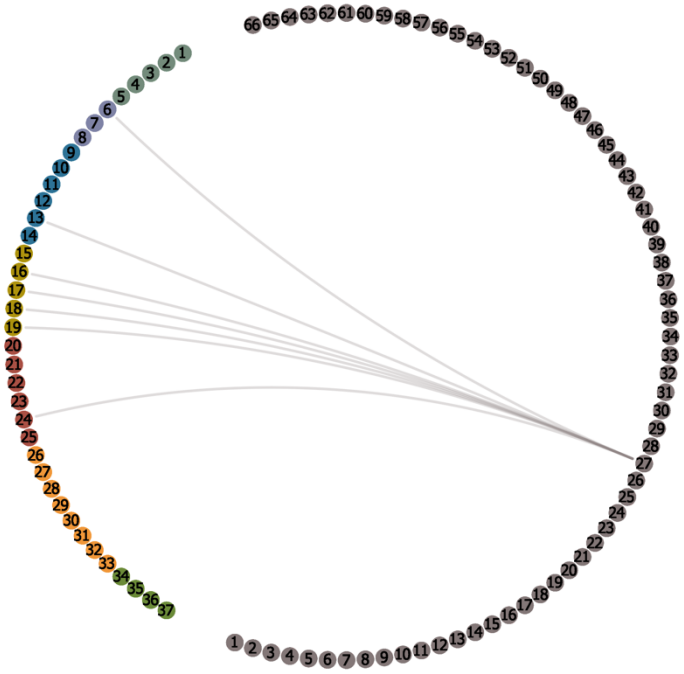


Figure 9 : Karen's Directory Printer



Le diagramme, devenu interactif, pourrait ainsi servir d'instrument de veille et d'aide à la décision, s'il est mis à la disposition du public. En effet, tout·e praticien·n·e, ou personne intéressée par l'archivage des documents numériques pourrait sélectionner une tâche ou une fonctionnalité et avoir accès à une liste d'outils proposant cette fonctionnalité.

À partir d'une liste d'outils, il serait aussi possible pour un·e archiviste de prendre connaissance de chacune des fonctionnalités proposées par un outil spécifique et de savoir ainsi si l'outil en question répond bel et bien à ses besoins archivistiques.

En outre, par le biais d'un formulaire disponible en ligne, dont les champs correspondraient aux catégories du tableau synoptique et à la liste des fonctionnalités normalisées, de nouveaux outils pourraient, à terme, être ajoutés par tous et toutes. Cela permettrait d'atteindre une plus grande exhaustivité tout en maintenant à jour la liste actuelle des outils, au gré des évolutions technologiques.

Enfin, si de nouvelles fonctionnalités techniques devaient être créées, il serait possible de les ajouter progressivement dans les options du formulaire.

3.3 Études de cas

3.3.1 Étude de cas n° 1 : Archifiltre

Développé « au sein d'une start-up d'État faisant partie de la Fabrique numérique des ministères sociaux », par une équipe qui associe « les archivistes des ministères sociaux et des profils techniques complémentaires (Développeur, UX designer, Product Manager, etc) » (Archifiltre - Fabrique des ministères sociaux 2022), Archifiltre est un outil récent (2018), libre et gratuit (le code source est public et accessible).

Fiche de présentation

1.	Nom	Archifiltre
2.	Développeur	Fabrique numérique des ministères sociaux (État Français)
3.	Site internet	https://archifiltre.fabrique.social.gouv.fr/ (consulté le 7.1.2022)
4.	Public cible souhaité	« Les professionnel·le·s de l'information (dont environ 5000 archivistes exerçant dans le secteur public) ont pour mission de traiter cette production exponentielle. Sans outil adapté, ils/elles étaient dans l'incapacité d'appréhender et donc de traiter ces volumes d'information pourtant potentiellement stratégique. Leur intervention consiste à évaluer la pertinence

		de l'information et ainsi trier, améliorer des arborescences de fichiers mais également accompagner les agents dans la réorganisation ou encore la recherche de leurs documents. » (beta.gouv [s.d.])
5.	Présentation par le développeur	« 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). » (SocialGouv 2022)
6.	Date (dernière mise à jour)	v3.2.2 – 22.09.2021 (état au 7.1.2022)
7.	Écrit en	JavaScript
8.	Configuration requise	Linux, MacOS, Windows 64 bits, Windows 32 bits
9.	Distribution	Libre, licence MIT
10.	Respect de la politique des données	« Archifiltre n'exploite pas les données personnelles des utilisateur·trice·s. Cependant, nous utilisons Matomo, un outil open-source pour avoir des statistiques d'utilisation sur notre outil. » (Archifiltre - Fabrique des ministères sociaux 2022)
11.	Interopérabilité	Permet l'exportation de métadonnées en XML selon le Standard d'échange de données pour l'archivage (SEDA) 2.1[7].
12.	Diversité des formats pris en charge	« Archifiltre prend en compte toutes les extensions de fichiers et exclut juste les fichiers cachés et système de l'analyse. De plus, la plupart des extensions de fichiers auront un code couleur dans l'arborescence, mais celles qui ne peuvent pas être catégorisées seront affichées en gris. » (SocialGouv 2022)
13.	Taille maximum	en théorie, pas de maximum, mais les temps de chargement peuvent être extrêmement longs avec de grandes volumétries.
14.	Langue(s)	français, anglais, allemand
15.	Assistance	Archifiltre propose un espace de co-création https://archifiltre.fabrique.social.gouv.fr/co-construction (consulté le 10.01.2022) « Les choix de développement sont basés sur les

[7] https://francearchives.fr/seda/ (consulté le 9.1.2022)

	retours et besoins formulés par les utilisateur·trice·s lors des openlabs ou auprès des ambassadeur·drice·s Archifiltre. » (Archifiltre - Fabrique des ministères sociaux 2022). Les développeurs organisent ainsi des « Openlab », présentés comme un « moment d'échanges » entre l'équipe de projet et les utilisateurs·trice·s, des ateliers de co-construction pour discuter des fonctionnalités et de l'interface de l'outil.
16. Manuel d'utilisation	Wiki bien documenté sur *GitHub* : https://github.com/SocialGouv/archifiltre/wiki/Wiki-Archifiltre (consulté le 7.1.2022)

Rapport de test

17. Installation et prise en main	Une fois le fichier téléchargé, aucune installation nécessaire, en ouvrant le fichier, Archifiltre s'exécute immédiatement. Beaucoup d'efforts mis dans l'ergonomie et la facilité d'utilisation : codes couleur lisibles, interface épurée et facilement compréhensible. S'adresse tout à fait à un public « non technique ».

18. Fonctionnalités promises sur le site (Archifiltre - Fabrique des ministères sociaux 2022)
    a. « Appréhender des arborescences » ; « visualisez vos dossier et fichiers selon différentes pondérations et modes de classement »

    Avec un *drag and drop* du dossier à analyser, le logiciel charge une représentation en « stalactite » de l'arborescence : le dossier de base prenant toute la largeur de la fenêtre, puis chaque niveau de profondeur se développe vers le bas, chaque dossier étant pondéré par taille (plus un dossier est lourd, plus il sera large). En fin de chaîne, on trouvera les documents, représentés par les codes couleur classiques de la suite word (vert pour Tableurs, rouge pour Présentations, bleu pour Documents et violet pour Images, les dossiers sont représentés en jaune). Il est également possible de pondérer par nombre de fichiers. Aussi, il est possible de choisir une coloration non par type de dossiers mais par date d'ouverture afin de visualiser quels documents ont été plus ou moins récemment ouverts.

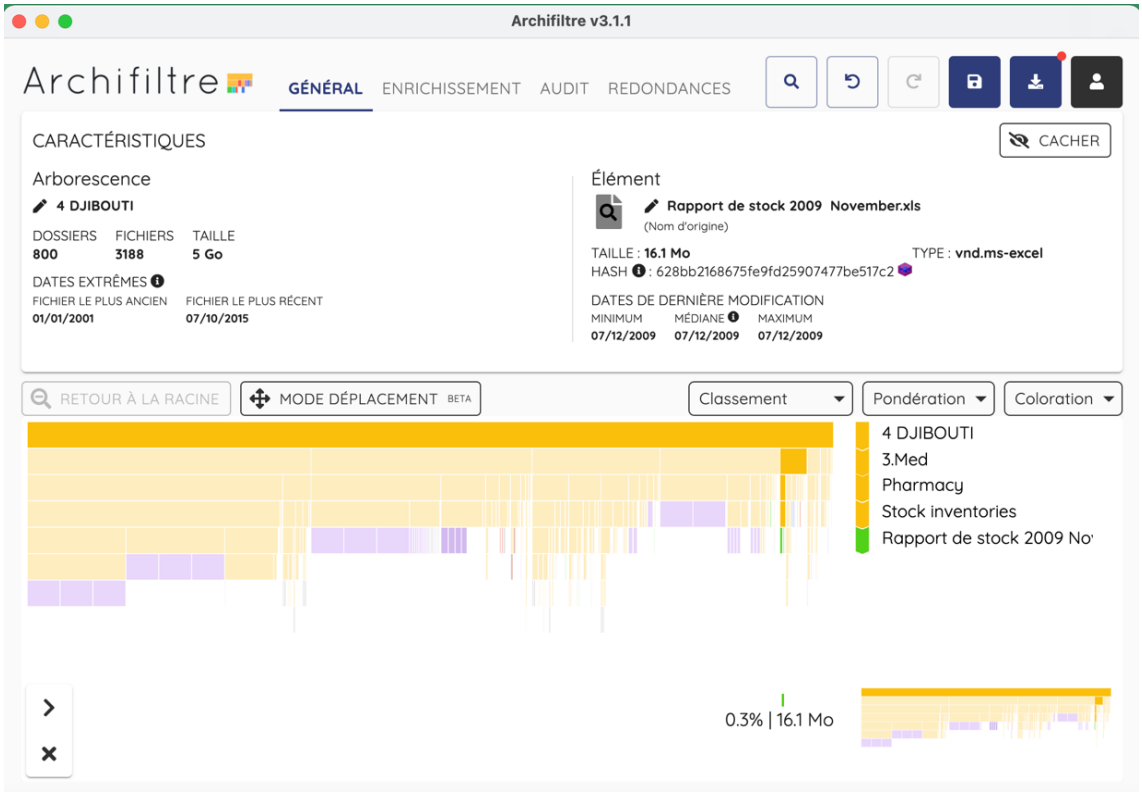
    Lorsqu'on sélectionne un fichier (« Rapport de Stock 2009 novembre.xls », Figure 10), Archifiltre affiche dans les fenêtres du haut à gauche les caractéristiques principales pour l'arborescence (nombre de dossiers, de fichiers, taille totale, dates extrêmes de modification) et à droite les métadonnées du fichier sélectionné (taille, empreinte générée par l'algorithme MD5, format et date de modification).

Cette fonctionnalité est très utile lorsqu'on appréhende un « vrac » numérique sans connaissances préalables : on se fait très rapidement une idée de la taille et de la structure du fonds, des indicateurs fondamentaux pour une *broad evaluation* (Belovari 2017).

Ici, cela nous permet de voir la structure grossière du jeu de données : une division par département (Médical, Logistique, RH, etc) et où se trouvent les fichiers les plus lourds, sur lesquels on peut – par exemple – prioriser nos efforts d'évaluation pour gagner un maximum de place de stockage.

Un « mode déplacement » (en version bêta) permet de remanier l'arborescence en déplaçant les dossiers au sein même de la visualisation. Archifiltre ne modifie cependant jamais directement les fichiers : il exportera un script qui peut ensuite être utilisé pour faire les changements souhaités (renommage, dédoublonnage, changements de structure).

Figure 1010 : Visualisation d'arborescence « en stalactite » dans Archifiltre



b. « Enrichir des métadonnées »

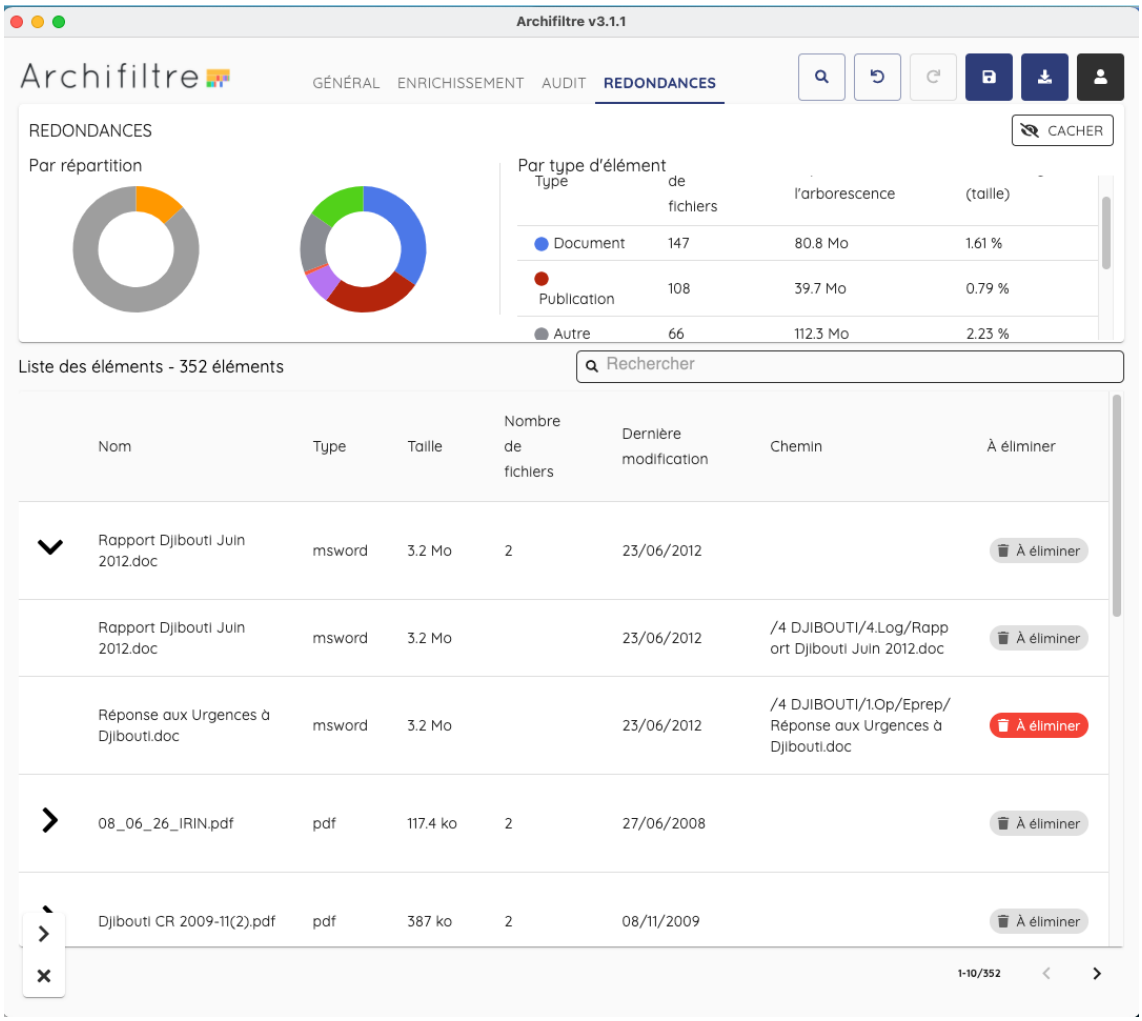
Dans l'onglet « Enrichissement », lorsque l'on sélectionne un fichier, apparaissent deux fenêtres supplémentaires : l'une « Description » où l'on pourra annoter le dossier (par exemple si on ne sait pas encore s'il doit être détruit ou déplacé ou si plusieurs archivistes travaillent sur le même fonds) ;

et une fenêtre « Tag » permettant de… taguer les documents, soit avec une étiquette : « à éliminer », soit en créant une ou des catégories *ad hoc* (« à revoir », ou « RH » par exemple) pour pouvoir ensuite filtrer par ces différentes catégories.

c. « Trouvez automatiquement les redondances de vos répertoires à partir du calcul d'empreintes »

Archifiltre dispose d'une fonctionnalité de Dédoublonnage – le terme choisi par les développeurs est « redondances ». Grâce à l'empreinte générée par MD5, le logiciel fournit une liste des doublons et quelques statistiques à leur sujet : quels formats sont à double, quel pourcentage du jeu de données cela représente, etc. (Figure). Il convient ici de signaler que le calcul des empreintes peut vite prendre beaucoup de temps lorsqu'on manie des échantillons d'une dizaine de GB par exemple. L'utilisateur doit ensuite sélectionner pour chaque fichier lequel il veut taguer « à éliminer » ; ces fichiers apparaîtront alors en rouge dans la visualisation par stalactites et il sera possible de générer un bordereau d'élimination contenant les fichiers ainsi tagués.

Figure 11 : Interface Dédoublonnage dans Archifiltre



d. « Mener une opération d'audit »

Un onglet « Audit » permet d'accéder à une fenêtre d'audit de l'arborescence où l'on trouvera (assez peu) d'informations : nombre de fichiers, niveaux de profondeur, et répartition par fichiers. L'intérêt de l'audit dans Archifiltre se trouve beaucoup plus dans la génération d'un rapport d'audit qui compile un certain nombre de « chiffres clés » de l'arborescence : nombre de dossiers/fichiers, dates extrêmes, mais aussi nom du fichier le plus long (les chemins trop longs peuvent être très problématiques car ils deviennent illisibles par Windows), ou encore le top 5 des répertoires les plus anciens/volumineux. Ce genre de rapports synthétiques peut être très utile pour faire naître une prise de conscience au sein des équipes ou de la hiérarchie quant à l'importance de la gestion quotidienne des fichiers numériques.

e. « Traitez en masse vos répertoires grâce aux exports Archifiltre (transferts, élimination, réorganisation) »

Archifiltre permet d'exporter des rapports de récolement des fichiers et dossiers analysés avec les métadonnées suivantes : chemin ; longueur du chemin ; nom ; extension ; poids (octet) ; date de première modification ; date de dernière modification ; nouveau chemin ; nouveau nom [si le fichier a été renommé] ; description ; fichier/répertoire ; profondeur ; nombre de fichiers type [pour les dossiers] ; empreinte (MD5) ; redondance [oui/non]].

19. Conclusion/Analyse

Une des grandes différences entre l'évaluation archivistique de documents papiers et de documents numériques, c'est que les contenus numériques sont souvent « invisibles » pour l'archiviste : « *Even just "knowing" what is contained in a digital collection can be challenging. In consequence, archivists have begun to call for simple digital approaches and tools*. » (Belovari 2017, p. 57). L'outil Archifiltre est un parfait exemple d'une approche « simple », accessible aux praticien·ne·s sans grandes connaissances techniques préalables. Dans l'exemple du jeu de données de *Médecins Sans Frontières*, provenant d'une mission de terrain à laquelle l'archiviste du siège n'a que peu accès voire de connaissances, Archifiltre permet d'« appréhender » – pour reprendre la terminologie utilisée par ses développeurs – le fonds en question, sa structure et, dans le cas présent, observer que le fonds respecte une structure semblable à celle utilisée par le siège.

3.3.2 Étude de cas n° 2 : DROID (Digital Record Object Identification)

DROID a été développé par les Archives nationales du Royaume-Uni pour l'identification automatisée par lots de formats de fichiers (*automated batch identification of file formats*). L'outil est prévu pour utiliser la base de données PRONOM[8] et ainsi répondre aux exigences fondamentales de tout dépôt numérique de pouvoir identifier précisément les formats des objets stockés et relier cette identification à un registre central d'informations techniques (The National Archives UK 2021a).

Fiche de présentation

1.	Nom	DROID (Digital Record Object Identification)
2.	Développeur	The National Archives (UK)
3.	Site internet	https://www.nationalarchives.gov.uk/inform

8 https://www.nationalarchives.gov.uk/PRONOM/BasicSearch/proBasicSearch.aspx?status=new (consulté le 9.1.2022)

		ation-management/manage-information/preserving-digital-records/droid/ (consulté le 07.01.2022)
4.	Public cible souhaité	« 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" (The National Archives UK 2020b, p. 3)
5.	Présentation par le développeur	« 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. » (The National Archives UK 2021a)
6.	Date (dernière mise à jour)	v. 6.5.1 – 01.05.2020
7.	Écrit en	JAVA
8.	Configuration requise	« build and tested on : Linux CentOS (Red Hat); Microsoft Windows 10 (64 bit); Raspbian; Mac OSX (Mojave); Mac OSX (Sierra); Linux Ubuntu Desktop » (The National Archives UK 2021a)
9.	Distribution	Libre, BSD Licence
10.	Respect de la politique des données	[non renseigné]
11.	Interopérabilité	[non renseigné]
12.	Diversité des formats pris en charge	Se base sur la base de données PRONOM en constante augmentation, 1400 formats (état mai 2020 (The National Archives UK 2020b, p. 3))
13.	Taille maximum	64 KB
14.	Langue(s)	anglais
15.	Assistance	Google Groups discussion page (https://groups.google.com/g/droid-list consulté le 07.01.2022) ; renvoient à une adresse E-mail en cas de questions : PRONOM@nationalarchives.gov.uk
16.	Manuel d'utilisation	oui, disponible ici : https://www.nationalarchives.gov.uk/documents/information-management/droid-user-guide.pdf (consulté le 07.01.2022)

Rapport de test

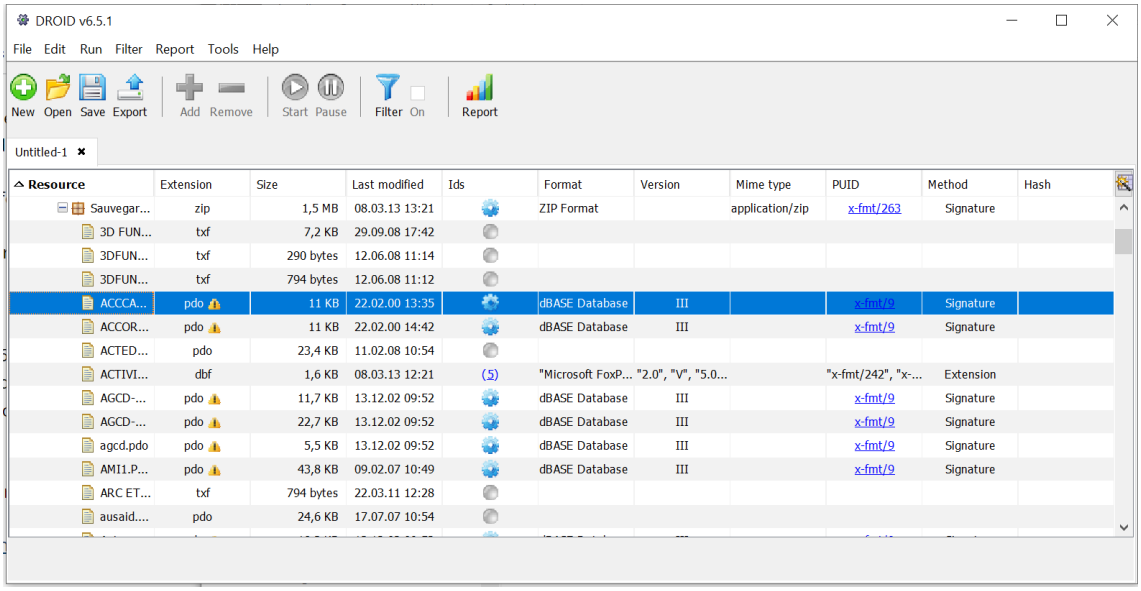
17. Installation et prise en main — Téléchargement d'un fichier .zip à extraire et installer dans un dossier choisi puis lancer en ouvrant le fichier « droid.bat » sur Microsoft ou « droid.sh » sur Mac OS ou Linux. Une interface brute et simple, ne laisse pas transparaître toutes les possibilités qu'offre le logiciel.

18. Fonctionnalités promises (The National Archives UK 2020b)

    a. « accurate file identification, even if the file extension is wrong of missing » ; « … also extracts other information about files it scans such as, file size, last modified, date and file path »

    Les documents ou fichiers à analyser peuvent être *drag and drop* dans l'interface ; les dossiers apparaissent ensuite dans l'interface et on peut naviguer dans l'arborescence. L'analyse est lancée en sélectionnant « Start ». Une fois l'analyse terminée, DROID indique toutes les erreurs de formats (inconnus ? erronés ? mal identifiés ?). Les PUID sont des identifiants uniques de PRONOM, en cliquant dessus, nous sommes redirigés vers la page de PRONOM du format en question. DROID a trois méthodes pour identifier les formats, plus ou moins fiables : *extension* (DROID s'est simplement basé sur l'extension du fichier – peu fiable) ; *signature* (DROID a détecté des « motifs » (*patterns*) dans les séquences d'octets propres à chaque format et méthode) ; *container* (DROID va vérifier dans le document même s'il y a des signatures propres – il s'agit de la méthode la plus précise).(The National Archives UK 2020b, p. 10).

Figure 12 : Interface DROID

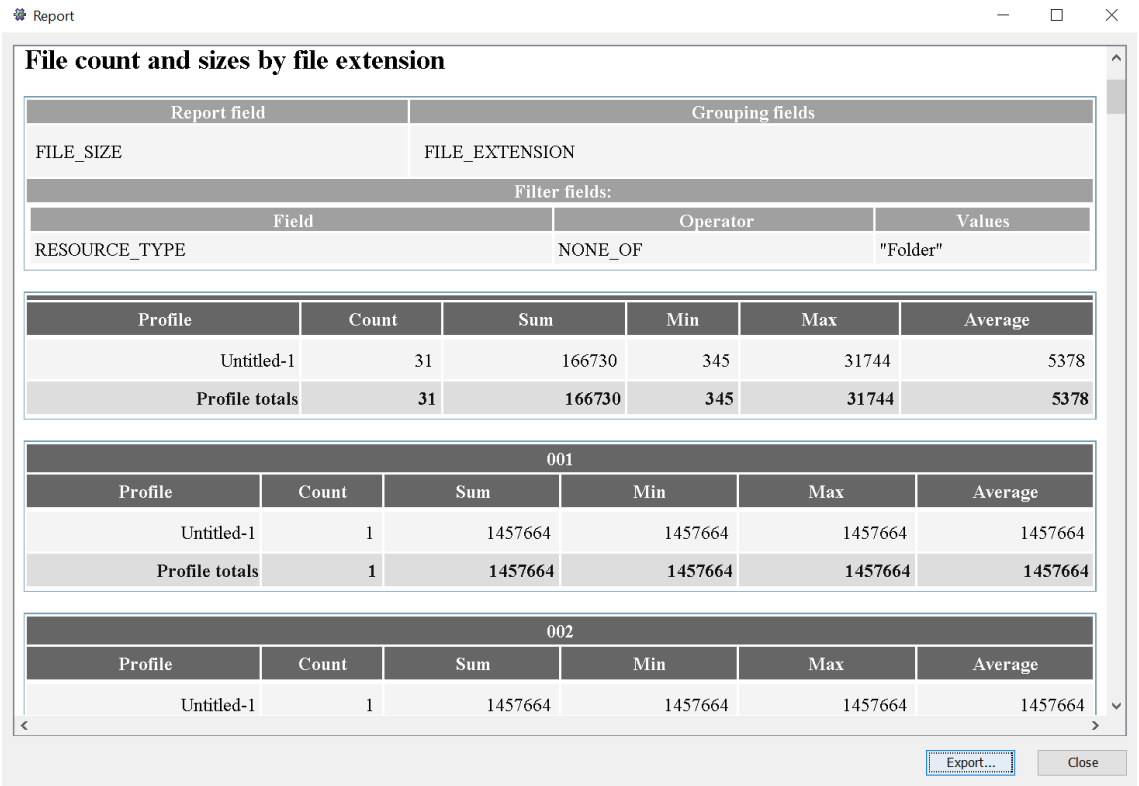


b. « …information is presented in a profile which can be analyzed on screen in the DROID Graphical User Interface (GUI) using filtering »

Une fois les fichiers à analyser importés, DROID permet de filtrer le résultat selon une série de critères (*File name ; File size ; File extension ; Last modified*, etc), une fonctionnalité très utile surtout pour les grands jeux de données. On peut par exemple créer une liste contenant tous les fichiers non-lisibles en choisissant le critère « *Job status* » et les valeurs « *Not found* » ; « *access denied* » et « *Error* ». Ou encore choisir le filtre « *Exstension missmatch* » avec critère « *true* » pour accéder et analyser les erreurs d'extensions. Divers filtres peuvent être appliqués (« plus grand que X » *et* « du format Y » par exemple).

Tous les résultats de recherche peuvent ensuite être exportés sous forme de divers rapports, par exemple un « *comprehensive report* », un rapport synthétique, ici (Figure) nous avons généré un rapport par extension. Ces rapports peuvent être exportés sous divers formats.

Figure 13 : Rapport synthétique DROID



c. or exported to a CSV file

Une manière d'analyser les résultats est d'exporter un fichier CSV qui peut ensuite être analysé dans un tableur. DROID extrait les métadonnées suivantes des fichiers et dossiers :

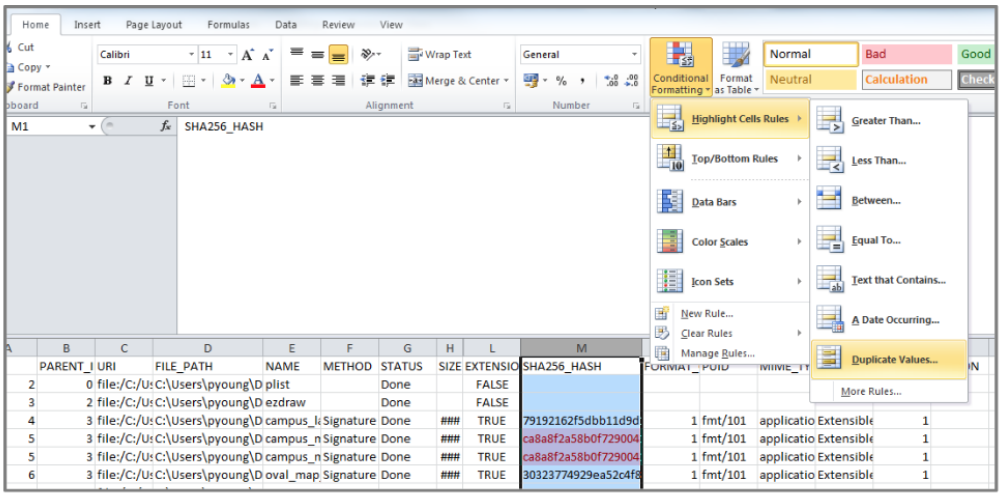
ID ; parent ID ; unique resource identifier (URI) ; file path; filename; identification method (signature, container signature or extension); status ; file size ; type ; file extension ; last modified date ; extension mismatch warning ; hash ; file format count ; PRONOM unique identifier for the file format (PUID) ; mime-type ; file format name ; file format version.

Ce rapport de récolement, semblable à celui proposé par Archifiltre, propose quelques métadonnées un peu plus précises sur les formats de fichiers. On notera que le manuel d'utilisateur suggère une manière très « manuelle » de dédoublonner les fichiers, en important le fichier CSV dans excel puis en sélectionnant la colonne des empreintes (*hash*) et enfin en surlignant les valeurs à double (Figure).

Figure 14 : Dédoublonnage dans DROID

Detecting duplicates

An easy way to see if you have duplicate files using Excel is to highlight the HASH column and select the Home section - Conditional Formatting - Highlight Cells Rules - Duplicate Values. You can then choose a colour which Excel will use to highlight each cell which has duplicate values.



(The National Archives UK 2020b, p. 12)

19. Conclusion/Analyse

DROID fait partie des outils de référence pour l'identification de format, enjeu central dans la préservation numérique, pour garantir l'accessibilité et se prémunir contre l'obsolescence des formats.

3.3.3 Étude de cas n°3 : Karen’s Directory Printer

Karen’s Directory Printer est un petit outil d’assistance (désigné de *Power tool* par les développeurs) permettant de créer des répertoires personnalisés de fichiers.

Les éléments entre guillemets sont tirés du site internet du logiciel (Karenware 2022)

Fiche de présentation

1.	Nom	Karen’s Directory Printer
2.	Développeur	Karen’s Power Tools
3.	Site internet	https://www.karenware.com/powertools/karens-directory-printer (consulté le 08.01.2022)
4.	Public cible souhaité	[non précisé]
5.	Présentation par le développeur	« 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. » (Karenware 2022)
6.	Date (dernière mise à jour)	v. 5.4.4, 25.05.2020
7.	Écrit en	[non renseigné]
8.	Configuration requise	Windows
9.	Distribution	libre
10.	Respect de la politique des données	[non renseigné]
11.	Interopérabilité	[non renseigné]
12.	Diversité des formats pris en charge	[non renseigné]
13.	Taille maximum	[non renseigné]
14.	Langue(s)	anglais
15.	Assistance	https://helpdesk.karenware.com/ (consulté le 8.1.2022) page désactivée, *GitHub* en construction
16.	Manuel d’utilisation	non

Rapport de test

17.	Installation et prise en main	fichier .exe de 1,8 MB à télécharger et exécuter. Une seule fenêtre avec quatre onglets, les

informations sont denses, mais sans fioritures

18. Fonctionnalités promises
    a. « Select a drive or directory and […] print the name of every file…"
    b. "… along with the file's size, …"
    c. "… date and time of latest modification, …"
    d. "… and attributes (Read-Only, Hidden, System and Archives)"
    e. "… the list can be sorted by name, size, date created, date last modified, or date of last access."

    Pour créer un répertoire personnalisé, l'utilisateur·trice sélectionne le dossier en question puis choisit parmi les métadonnées proposées : les classiques « dates de création », « taille » (Figure), mais il est également possible de personnaliser les recherches et par exemple de filtrer par formats de sons, images ou fichiers exécutables « communs » (Figure).

Figure 15 : Interface principale de Karen's Directory Printer

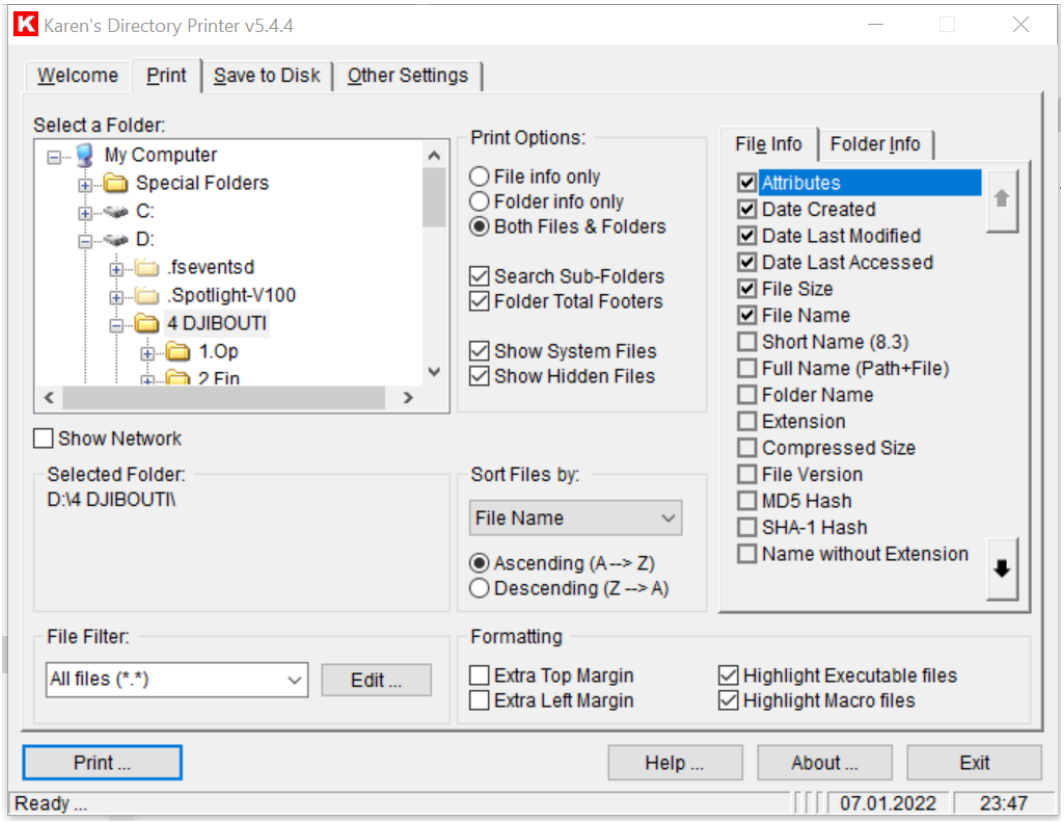
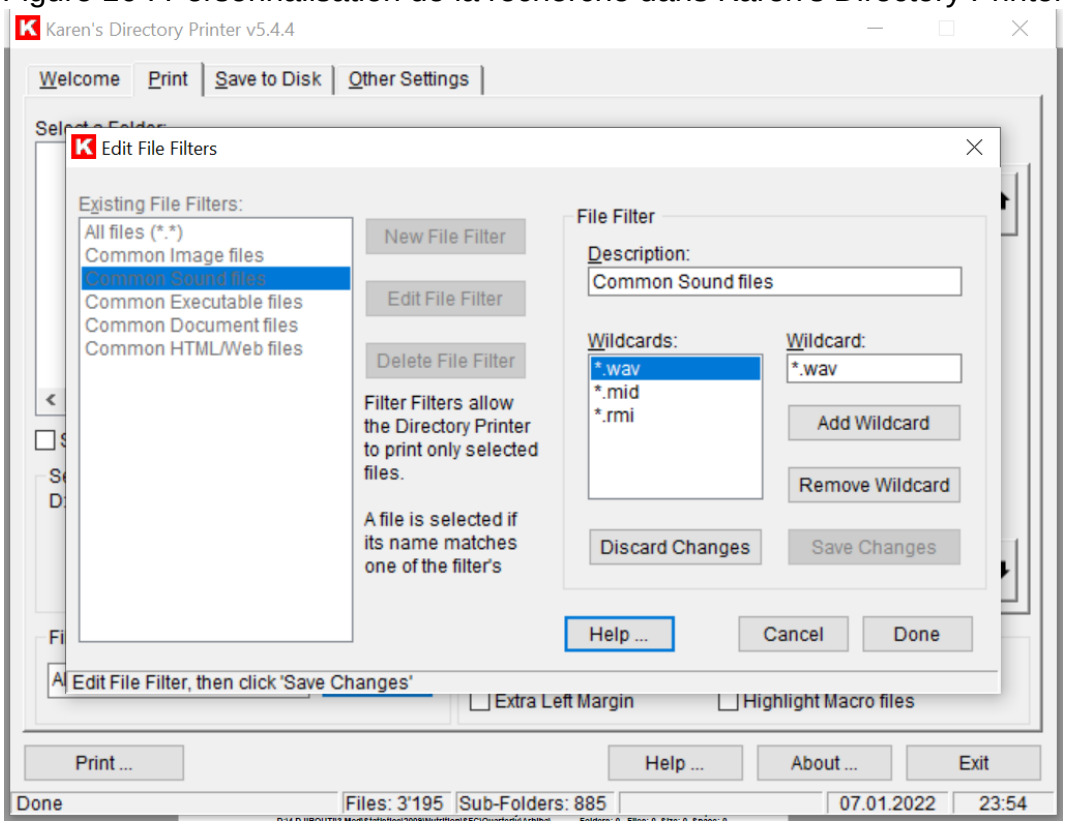


Figure 16 : Personnalisation de la recherche dans Karen's Directory Printer



19. Conclusion/Analyse

Certes, l'outil Karen's Directory Printer peut sembler bien loin de l'automatisation, voire carrément vétuste. Cependant, le choix de cet outil est assumé car il représente bien une famille d'outils : légers, gratuits, faciles d'utilisation et mono-tâche avec une approche pragmatique. En effet, Karen's Directory Printer a été créé en 1997 – soit une éternité dans le domaine numérique – et pourtant il est encore mis à jour, preuve qu'il y a toujours une demande pour des outils se trouvant à l'opposé de l'*autoclassification* (National Archives and Records Administration 2014) dans le spectre de l'automatisation.

4. Discussion

4.1 Discussion des résultats

4.1.1 Du soutien à la réalisation d'une tâche archivistique

Comme nous l'avons explicité dans la méthodologie et dans les résultats de notre projet de recherche, nous avons décidé d'attribuer une valeur positive (qui prend alors la forme d'un « 1 » dans le tableau à double entrée et d'une « liaison » dans le diagramme de membrures) si une fonctionnalité informatique « soutenait » la réalisation d'une tâche archivistique. En d'autres mots : si une fonctionnalité permettait d'*aider* l'archiviste dans la réalisation d'une activité spécifique. Cet encodage systématique perdait alors en finesse ce qu'il avait gagné en efficacité. En effet, comment distinguer le *niveau* de soutien ou de réalisation d'une tâche via une ou plusieurs fonctionnalités à l'aide d'un codage binaire ?

Comme on peut le constater dans le tableau, il existe parfois une fonctionnalité dont l'intitulé correspond peu ou prou à la définition d'une tâche archivistique : c'est par exemple le cas pour la tâche n° 31, « Convertir les formats de fichiers » et la fonctionnalité n° 19, « Conversion de formats ». Le nombre de « 1 » qui figurera dans le tableau sera alors probablement faible, et seule(s) une ou quelques liaisons reliront la fonctionnalité et la tâche en question. Sur le large spectre de l'automatisation, on sera ainsi amené à considérer que cette tâche, de nature très *technique* et *informatique*, est pleinement réalisée, et le plus souvent de manière automatique.

À l'inverse, et à titre d'exemple, il n'existe pas de fonctionnalité(s) éponyme(s) pour les tâches n° 22, « Analyser le contenu des documents », et n° 25, « Indexer ». Comme plusieurs tâches de nature très *intellectuelle* qui figurent dans le tableau, ces activités spécifiques sont extrêmement complexes et ne peuvent être réalisées par le biais d'une *seule* fonctionnalité. C'est alors bien une *combinaison multiple* de (micro)fonctionnalités qui permettra à l'archiviste de réaliser *in fine* la tâche en question. Raison pour laquelle il existe de nombreux « 1 » dans le tableau et beaucoup de « liaisons » dans le diagramme de membrures pour certaines tâches archivistiques très intellectuelles (qui pourraient être encore décomposées en une suite d'activités très spécifiques) qui sont *soutenues* par plusieurs fonctionnalités distinctes.

Ainsi, il serait méthodologiquement erroné de considérer que les tâches réunissant le plus de « 1 » sont plus automatisées que celles ne comptabilisant qu'un nombre restreint de liaisons. Il est effectivement nécessaire de prendre en compte l'adéquation entre une tâche archivistique et une fonctionnalité. Potentiellement, plus une tâche est

(directement) automatisée, moins elle nécessite de fonctionnalités distinctes ; à l'inverse, s'il n'existe pas de fonctionnalité réalisant (exactement) une tâche, la bonne exécution de cette activité s'appuiera sur une combinaison de fonctionnalités différentes et complémentaires.

4.1.2 Des tâches techniques, intellectuelles et humaines

Les remarques qui précèdent, relatives à l'adéquation potentielle entre une fonctionnalité et une tâche, permettent entre autres de proposer une répartition des tâches selon des typologies distinctes. Après analyse du tableau et du diagramme de membrures, trois familles semblent se distinguer : les deux premières, déjà évoquées, regroupent les tâches techniques d'une part et les tâches intellectuelles d'autre part ; la troisième réunit des tâches que l'on pourrait qualifier d'« humaines ».

Les tâches dites « techniques » sont celles qui paraissent le plus soutenues, ou réalisées, par des fonctionnalités informatiques spécifiques (voir ci-dessus) : il s'agit principalement des tâches qui font partie de la fonction archivistique « Préservation ». Cela consiste surtout à manipuler de grandes quantités de documents et de données dans le but d'en préserver l'intégrité et de garantir leur exploitabilité future.

Plusieurs tâches peuvent être qualifiées d'« intellectuelles » dans la mesure où il est nécessaire de mener une réflexion approfondie en amont ou durant le traitement archivistique. Si ces tâches ne sont pas totalement automatisées, leur plein accomplissement peut néanmoins s'appuyer sur un ensemble de fonctionnalités qui aideront l'archiviste au quotidien, en particulier dans la prise de connaissance du « contenu » des documents. Il s'agit alors le plus souvent des tâches qui font partie des fonctions « Évaluation », « Classification » et « Description ».

La troisième « famille » concerne les tâches dites « humaines », pour lesquelles il n'existe à l'heure actuelle pas ou peu de soutien de type informatique : ces tâches sont actuellement peu automatisées, et leur potentiel d'automatisation paraît relativement faible. Dans le tableau, cela se traduit par une succession ininterrompue de « 0 » et par une absence complète de liaisons dans le diagramme de membrures. Les tâches suivantes peuvent, par exemple, être qualifiées d'« humaines » : « Établir la relation personnelle avec les donateurs » ; « Identifier les usagers et leurs besoins » ; « Former les usagers » ; « Développer la clientèle ». Comme on peut le constater, ces tâches appartiennent principalement aux fonctions archivistiques : « Accroissement » (et tout particulièrement la sous-fonction « Acquisition » pour ce qui concerne les archives non-institutionnelles/privées) et « Diffusion ». Situées aux deux extrémités symboliques du processus archivistique (acquérir des archives auprès de personnes ou d'institutions non

soumises à une obligation légale de dépôt ; mise en valeur des archives conservées et mise à disposition des documents aux utilisateurs finaux), ces deux fonctions nécessitent des contacts humains et requièrent des relations interpersonnelles (bien que des outils informatiques propres à alléger le travail quotidien des archivistes soient progressivement développés dans ce domaine-là également : des *chatbots* pour communiquer avec les utilisateurs·trices ; des *Massive Open Online Course (MOOC)* pour former les usagers·ères ; ou des instruments d'analyse statistiques comme *Google Analytic*s pour identifier les besoins et les demandes des chercheurs·ses).

4.1.3 De l'utilité de certaines fonctionnalités : le cas du *NLP*

Parmi les fonctionnalités au plus fort potentiel pour l'archiviste – ou disons : celles qui lui permettent d'effectuer plus efficacement une tâche archivistique – figure en particulier le *Natural Language Processing* (*NLP*, soit, en français, le *Traitement automatique du langage naturel*). Il s'agit d'un domaine d'étude multidisciplinaire situé au croisement de la linguistique et de l'intelligence artificielle utilisé pour l'analyse automatique et la représentation du langage humain (Young et al. 2017) nécessitant deux étapes successives : le *pre-processing* pour préparer les données puis le développement d'un algorithme soit *rules-based*, basé donc sur des règles préétablies, soit *machine learning-based system*, qui utilise des méthodes d'apprentissage et des échantillons d'entraînement pour apprendre progressivement.

Cet ensemble de méthodes regroupées sous le terme *NLP* se révèle central pour l'archiviste puisqu'il permet d'extraite du texte, de l'analyser et de prendre connaissance du contenu des documents de manière extrêmement efficace. Plusieurs fonctionnalités listées dans notre tableau font partie du *NLP* (dont « *Topic modelling* », « *Sensitivity Review* » ou « Reconnaissance d'entités nommées ») et soutiennent la réalisation d'un grand nombre de tâches, notamment dans le domaine de l'évaluation archivistique. Si l'évaluation « large » – appelée « *broad digital appraisal* » par Belovari (2017, p. 57) –, visant à identifier et supprimer les doublons, les fichiers vides, inutiles et temporaires, de sorte à réduire la taille et la complexité des documents numériques, peut être réalisée via le dédoublonnage ou la production de rapports synthétiques par exemple, les fonctionnalités du *NLP* permettent, quant à elles, de procéder à une évaluation beaucoup plus fine et d'attribuer une valeur archivistique à un ensemble de documents – une évaluation intitulée « *in-depth digital appraisal* » (Belovari 2017, p. 57). Ces fonctionnalités sont enfin particulièrement efficaces lorsqu'il s'agit de décrire les documents (fonction archivistique de « Description ») puisque l'on peut ainsi prendre connaissance de manière automatique et pour de grands volumes de données du contenu des documents (Hooland et Coeckelbergs l'ont par exemple démontré de

manière convaincante en 2018 en utilisant du *Topic modeling* sur des lots de documents numériques de la Commission européenne).

Toutefois, si cette technologie paraît très prometteuse dans le domaine des archives, de nombreuses adaptations seront encore nécessaires pour correspondre aux attentes et aux besoins des archivistes. Les outils proposant des fonctionnalités basées sur le *NLP* sont par exemple peu intuitifs, les interfaces sont fréquemment assez complexes – voire austères –, et il est souvent nécessaire de disposer de compétences informatiques relativement poussées pour profiter pleinement des avantages du *NLP* et en comprendre toutes les ressources. Dans ce domaine plus qu'ailleurs peut-être, il sera alors nécessaire de former des équipes interdisciplinaires (d'archivistes et de spécialistes *IT*), de parfaire la formation technologique et informatique des archivistes, et de mettre en place des collaborations interinstitutionnelles (car les différents services d'archives n'ont pas les ressources pour créer et entraîner des algorithmes complexes et tenir à jour des dictionnaires) afin de partager les outils, les compétences et les expériences.

4.1.4 Des outils provenant d'horizons très divers

Comme nous avons pu le constater lors de la collecte d'informations à propos des outils et des fonctionnalités, il existe un grand nombre de produits qui n'ont pas été initialement développés pour l'accomplissement de tâches archivistiques. Le traitement des archives numériques peut ainsi s'appuyer sur un ensemble d'outils développés et de technologies mises à profit dans des domaines très divers.

L'exemple le plus frappant est sans doute celui de la *criminalistique numérique* (*digital forensics*). Il existe en effet un grand nombre de logiciels utilisés par des informaticiens pour la résolution d'enquêtes criminelles, comme des bloqueurs d'écriture (afin de garantir l'intégrité des fichiers), des outils de récupération de données (pour identifier les fichiers cachés ou les éléments supprimés) ou des visualisations chronologiques et géographiques (pour mettre en évidence l'existence de réseaux par exemple). En 2010, un rapport soutenu par le CLIR (*Council on Library and Information Resources*, Washington D.C.) et intitulé significativement « Digital Forensics and Born-Digital Content in Cultural Heritage Collections » mettait d'ailleurs en avant les avantages que les institutions patrimoniales pouvaient retirer de l'utilisation de la méthodologie et des outils développés dans le domaine de la criminalistique informatique :

> *« Incorporating forensics methodology and tools into the archival workflow will enable digital archivists and curators to capture more information about the content and makeup of digital objects ; help repositories manage the data copied from disks more efficiently and in accordance with established standards ; reinforce the importance of documentation to all aspects of the curation cycle ; and give archivists, donors, and others the ability to preview the contents of both*

isolated storage media and complete computing systems to formulate acquisition and preservation strategies » (Kirschenbaum, Ovenden, Redwine 2010, p. 60)

Très prometteurs, ces outils (comme *Forensic ToolKit* d'Exterro ou *EnCase* d'Opentext) ne sont cependant pas à la portée de toutes les institutions (en termes de compétences techniques nécessaires à leur prise en main et de ressources financières pour les acquérir). En outre, les fonctionnalités offertes par ces outils doivent être adaptées aux domaines des archives, comme l'affirment avec force Kirschenbaum, Ovenden et Redwine (2010, p. 60) en conclusion de leur rapport : « the first and most compelling conclusion we have drawn from our research and conversations is that over the long term, digital forensics should not simply be imported and adopted in toto into manuscript archives and the broader cultural heritage and scholarly communities ». À titre d'exemple, signalons que les experts en criminalistique sont plus à la recherche de pièce(s) isolée(s) pouvant constituer une preuve d'un délit, tandis que les archivistes ont à manipuler de grandes quantités de données et que le besoin se fait surtout sentir dans le domaine de l'exploration et de la visualisation à grande échelle (Kirschenbaum, Ovenden et Redwine 2010, p. 60-61).

Les archivistes doivent donc se tenir au courant des développements technologiques mis en œuvre dans des domaines voisins, afin d'en tirer profit pour le traitement de volumineuses archives numériques. La curiosité et l'enthousiasme pour les technologies disponibles dans d'autres domaines doivent néanmoins toujours être encadrés par les principes archivistiques et déontologiques qui guident notre profession (respect des fonds, principe de provenance, qualité des archives, etc.).

4.2 Discussion de la méthode

4.2.1 Exhaustivité non garantie des outils et des fonctionnalités

La conséquence la plus dommageable de la méthode que nous avons utilisée pour lister les outils et les fonctionnalités est sans aucun doute l'impossibilité d'atteindre l'exhaustivité. Les outils ont effectivement été répertoriés sur la base de lectures réalisées suite à une recherche bibliographique qui s'est voulue la plus complète possible, mais qui n'est certainement pas exempte de lacunes. Focalisée sur des articles scientifiques traitant de la gestion des archives électroniques (*born-digital*), et plus particulièrement sur les possibilités offertes par l'automatisation des fonctions archivistiques, cette recherche bibliographique première (réalisée *via* les moteurs de recherches classiques et dans des bases de données thématiques) a été complétée par une collecte secondaire d'informations, par « effet boule de neige », soit l'étude des références bibliographiques et scientifiques données d'article en article. Si cette

approche *itérative* a le mérite d'atténuer ce biais, en prenant en compte de plus en plus de références, il est impossible, à ce stade de notre recherche, de garantir l'exhaustivité des outils référencés et des fonctionnalités normalisées.

Ce biais a des conséquences importantes puisqu'il influence négativement les conclusions de notre étude : des « lacunes » dans le domaine de l'automatisation (soit des tâches archivistiques peu soutenues par des fonctionnalités logicielles) peuvent avoir été identifiées de manière erronée et le praticien, qui utiliserait ce tableau comme aide à la décision, ne trouverait pas l'outil dont il a besoin pour gérer les documents numériques placés sous sa responsabilité. Le caractère *dynamique* de notre méthode permet néanmoins de répondre, partiellement du moins, à ce défaut : un outil peut être ajouté en tout temps au tableau synoptique, ses fonctionnalités normalisées (si elles ne figurent pas encore dans la liste) et le diagramme de membrures adapté en conséquence. Le biais résultant du manque d'exhaustivité est donc particulièrement important lorsqu'il s'agit de tirer des conclusions à un moment précis, de dresser un bilan à un instant *T* ; raison pour laquelle nous avons désiré également proposer, par cette étude, un cadre méthodologique et conceptuel amené à se développer et à être réutilisé dans le futur, sous une forme collaborative.

4.2.2 Disparité des informations mises à disposition

Lors de la collecte des informations relatives à chaque outil, nous avons pu constater une grande disparité dans le type, la quantité et la qualité de la documentation fournie par les développeurs. Sachant que les données présentées dans le tableau à double entrée et utilisées pour la réalisation du diagramme de membrures sont directement extraites des sites internet des développeurs, cette grande hétérogénéité peut engendrer des biais et des phénomènes de distorsion. On peut ainsi se méprendre sur l'étendue réelle des fonctionnalités offertes par un outil parce que ces dernières ne sont pas explicitement et exhaustivement présentées sur le site internet du développeur ; à l'inverse, les concepteurs d'un outil très spécifique peuvent fournir le détail complet de toutes les étapes de traitement et ainsi présenter une liste très importante et très précise de (micro)-fonctionnalités.

Ce phénomène recoupe, en tout cas partiellement, la distinction qui existe entre d'une part les outils *open source*, dont le code de programmation et les spécificités techniques sont souvent disponibles sur des plateformes collaboratives (comme *GitHub*), et d'autre part les logiciels *propriétaires*, pour lesquels il est difficile de saisir quelle est l'étendue exacte des fonctionnalités offertes. Des développeurs de grande envergure se livrent

une concurrence féroce sur le marché de la gestion de l'information numérique et, dans l'optique de convaincre de nouveaux clients, la présentation des outils prend alors la forme d'une campagne publicitaire : utilisation de slogans et de termes génériques (on s'en tient souvent aux fonctions archivistiques principales, comme l'accroissement ou la préservation), vulgarisation excessive, etc. En outre, ces développeurs proposent généralement des plateformes multiservices et intégrées (de la GED au SAE) permettant le traitement intégral du cycle de vie des documents, ou alors des solutions adaptées et sur mesures, en fonction des demandes de leurs clients.

Pour toutes ces raisons, il a été difficile, via la consultation des sites internet, d'intégrer les fonctionnalités détaillées des outils proposés par des géants comme *Opentext*, *Axiell*, ou *Everteam.* Ces outils mériteraient d'être testés à grande échelle avec des jeux de données distincts, et de faire l'objet d'études de cas individuelles.

4.2.3 Granularité inégale des fonctionnalités

De l'hétérogénéité des informations recueillies découle en partie la granularité inégale des fonctionnalités. Certains sites internet de concepteurs d'outils s'attardent longuement sur les composantes techniques d'une fonctionnalité (sans forcément mettre en avant le résultat final que peut en attendre un·e archiviste), tandis que d'autres se concentrent sur l'objectif final, soit la réalisation d'une certaine tâche archivistique via la mise en œuvre d'une série non précisée d'actions informatiques. La profondeur de la description varie ainsi considérablement (en fonction du nombre d'étapes différentes dont peut se composer une tâche), de même que le point de vue adopté : alors que certains développeurs privilégient le point de vue des utilisateurs finaux (qui ne s'intéressent peut-être pas aux coulisses techniques nécessaires à l'accomplissement d'une action), d'autres concepteurs détaillent les spécificités techniques de leur outil, sans même mentionner les possibilités ainsi offertes.

Un exemple illustre la difficulté sans cesse rencontrée : le dédoublonnage. Il s'agit d'une tâche essentielle dans le processus d'évaluation et de versement des archives puisqu'elle permet de ne conserver qu'une seule copie d'un fichier informatique rigoureusement identique et d'économiser ainsi l'espace de stockage (avec toutes les économies en termes de processus, de temps, d'argent et d'énergie que cela implique). Or pour réaliser cette tâche, il existe plusieurs possibilités, qui nécessitent chacune que l'outil réalise une suite de micro-actions : l'identification de doublons peut se faire grâce à un traitement de métadonnées (nom, taille, date, extension, etc., du fichier) via une extraction, une compilation et une comparaison de métadonnées ; ou par le biais des

« sommes de contrôle » (*checksum*), avec un calcul de la somme de contrôle de chaque fichier puis une comparaison inter fichiers. L'outil peut ensuite proposer un traitement de masse (comme la suppression de fichiers), l'enregistrement de l'action, ou encore la création d'un « lien dur » (*hard link*[9]) pour garantir la traçabilité et l'intégrité du fonds d'archives. La question à laquelle nous avons été confronté est donc la suivante : est-il plus pertinent d'opter pour une fonctionnalité généraliste et pratique (orientée vers l'objectif final que peut en attendre un·e archiviste) telle que « Dédoublonnage » (sans préciser exactement ce que le terme recoupe), ou vaut-il mieux décomposer cette fonctionnalité en ses multiples composantes techniques : « Calcul de sommes de contrôle », « Comparaison de sommes de contrôle », « Suppression automatisée des fichiers aux sommes de contrôle identiques », etc. ? Sachant en outre que ces composantes techniques entrent dans le processus d'autres fonctionnalités : c'est le cas des sommes de contrôle qui servent également à vérifier le bon déroulement d'un transfert et d'une migration de données (par le biais de la comparaison des sommes de contrôle avant et après ledit transfert, pour garantir que les données n'aient pas été altérées).

Dans le cadre de cette étude, nous avons généralement opté pour la première solution : des fonctionnalités aux intitulés généralistes et pratiques, qui reflètent les besoins archivistiques courants et peuvent être facilement compréhensibles par le plus grand nombre. Une harmonisation complète de la granularité des fonctionnalités est toutefois difficile à atteindre, raison pour laquelle il demeure un biais dans ce domaine : des fonctionnalités généralistes ou orientées action (comme « Contrôle de conformité des SIP », n° 15) voisinent encore, dans les résultats présentés, avec des fonctionnalités plus techniques et détaillées (comme « Extraction de métadonnées », n° 36) qui sont parfois à la source d'autres fonctionnalités, ou qui sont nécessaires à la bonne exécution de fonctionnalités plus globales.

[9] Belovari 2017 a notamment testé la fonctionnalité « *hardlinks* » proposée par *Tree Size Professional (TSP)* et définie par *TSP* de cette manière : « They map a path in the file system on your drive and allow access to the corresponding file. Normally, each hard link points to its own section on the disk, but it is also possible that several hard links point to the same section. TreeSize takes advantage of this fact when deduplicating. It points multiple existing files to the same data and frees up the disk space that was previously occupied by individual chunks of data. After deduplication, the paths in Windows Explorer are still as visible as before, but they all point to the same file. » (https://www.jam-software.com/treesize/hardlinks.shtml, consulté le 13.01.2021)

5. Conclusion

Le tableau à double entrée et le diagramme de membrures qui mettent en relation les tâches archivistiques et les fonctionnalités informatiques nous ont permis d'attirer l'attention sur les fonctions archivistiques qui bénéficient le plus des possibilités offertes par l'automatisation ; ou à l'inverse, celles qui sont actuellement peu soutenues par les outils existants.

Toutes les fonctions archivistiques ne sont effectivement pas soutenues de manière équivalente par des fonctionnalités informatiques. Alors que certaines semblent pouvoir s'appuyer sur un grand nombre d'outils et de fonctionnalités pour (semi-)automatiser leurs procédures, comme la *préservation* ; d'autres, à l'instar de *l'acquisition* ou de la *diffusion*, ont encore le plus souvent recours à des processus manuels ou disons : peu automatisés. Enfin, la fonction de *description* peut de plus en plus bénéficier de solutions et de technologies qui permettent de traiter plus efficacement de grandes quantités de données et de prendre connaissance rapidement du contenu des documents. L'identification de ces « lacunes » devrait ainsi permettre, dans le cadre d'une recherche future, de mettre sur pied des équipes interdisciplinaires d'archivistes et d'informaticien·ne·s pour développer des fonctionnalités et des outils propres à soutenir le travail quotidien d'un service d'archives dans les domaines encore peu ou pas automatisés jusqu'à présent.

Ce projet de recherche a également permis de référencer les outils actuellement disponibles avec les possibilités de traitement qu'ils offrent, afin que les praticien·ne·s puissent facilement prendre connaissance des solutions informatiques existantes. La nécessité pour les professionnel·le·s de disposer d'un catalogue d'outils et de services paraît d'autant plus urgente que le domaine est vaste et en constant développement. Comme on a pu le constater tout au long de cette recherche, les outils existants sont effectivement très hétérogènes : libres ou propriétaires, développés dans le cadre de projets universitaires collaboratifs, par des entreprises privées ou des particuliers, et proposant tour à tour des fonctionnalités très spécifiques ou des solutions logicielles complètes. Ces outils peuvent encore prendre la forme d'application web ou de programme à télécharger en local et se présenter comme des plateformes *multi-* et *micro-* services qui combinent et agrègent de plus petits logiciels informatiques. Si ce projet de recherche a pour ambition de dresser un bilan *actuel* des outils existants, il vise donc également à proposer une grille d'analyse et une méthodologie reproductibles et propres à suivre les développements technologiques futurs.

De fait, la grande majorité des outils listés et des fonctionnalités identifiées dans ce projet de recherche correspondent à la quatrième approche de l'automatisation définie par les *National Archives and Records Administration* aux États-Unis : « *Modular Re-usable Management Tools* ». Le caractère modulable, interopérable et *open source* de ces outils semble effectivement bien adapté au domaine des archives, et tout particulièrement pour le traitement des archives non-institutionnelles et des fonds privés (objets de nos études de cas), qui sont souvent remis aux institutions patrimoniales en une seule fois, sans qu'il ait été possible de procéder à une analyse des besoins et de mettre en place un calendrier de conservation *ad hoc* tout au long du cycle de vie des documents.

*

Si ce travail avait principalement pour objectif de s'interroger sur les potentialités informatiques de traitement documentaire, la question de la place de l'archiviste et de son rôle dans les années à venir n'a cessé de nous accompagner. Pour paraphraser le titre d'un article de Pascal Morisod, il s'agira effectivement de trouver la bonne formule et le juste équilibre pour un (heureux) ménage à trois « des archives, des machines *et* des hommes » (2018). Cette réflexion s'articule autour de plusieurs questions, à commencer par celle des compétences qu'il sera nécessaire d'acquérir et de développer dans le futur. Car si l'archivistique s'automatise de plus en plus, quels sont les savoirs et les compétences qu'il conviendra d'enseigner aux archivistes du futur ?

Trois directions, qui ne sont contradictoires qu'en apparence, semblent se dessiner. Renforcer plus encore les compétences sociales et humaines des archivistes puisqu'il existera sans doute toujours des tâches qui ne pourront être (totalement) automatisées (dans le domaine des archives privées par exemple, ou en amont de la création documentaire, dans l'interaction avec les services producteurs, et en aval, dans le contact avec les chercheur·e·s). Développer les compétences classiques (les « savoirs de l'historien et des spécialistes des sciences auxiliaires » dont parle Morisod en 2018) et les connaissances para-documentaires (comme le contexte législatif ou la sensibilité de l'opinion sur des sujets historiques ou en lien avec la protection des données personnelles). Acquérir des compétences informatiques et en science de l'ingénierie afin de mieux comprendre quelles sont les potentialités offertes par l'intelligence artificielle dans le domaine archivistique et afin de participer activement au développement de nouveaux outils et de fonctionnalités en collaboration avec des informaticien·ne·s spécialisé·e·s. Cette dernière compétence sera d'autant plus nécessaire que les systèmes basés sur des règles doivent être paramétrés par des opérateurs humains,

que les algorithmes d'intelligence artificielle doivent être entraînés sur des échantillons avec des résultats devant être validés et pondérés par des archivistes, et que seul·e·s les archivistes seront en mesure de dire quels seront les outils nécessaires à la bonne conservation et diffusion du patrimoine documentaire. Loin de rétrécir le champ d'action des archivistes, l'automatisation est bel et bien une opportunité à saisir !

Bibliographie

ALBERTS, Inge et VELLINO, Andre, 2013. The importance of context in the automatic classification of email as records of business value: A pilot study. *Proceedings of the American Society for Information Science and Technology* [en ligne]. 2013. Vol. 50, n° 1, pp. 1-2. [Consulté le 17 avril 2021]. DOI https://doi.org/10.1002/meet.14505001112. Disponible à l'adresse : https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/meet.14505001112

ARCHIFILTRE - FABRIQUE DES MINISTÈRES SOCIAUX, 2022. Archifiltre. [en ligne]. 2022. [Consulté le 7 janvier 2022]. Disponible à l'adresse : https://archifiltre.fabrique.social.gouv.fr/

Automatisation. *Encyclopædia Universalis*, 2022. [en ligne]. [Consulté le 27 mai 2021]. Disponible à l'adresse : https://www.universalis.fr/encyclopedie/automatisation/.

BAILEY, S., 2009. Forget electronic records management, it's automated records management that we desperetely need. *Records Management Journal*. 2009. Vol. 19, n° 2, pp. 91-97.

BANAT-BERGER, Françoise, 2012. Les fonctions de l'archivistique à l'ère du numérique. In : *Les chantiers du numérique. Dématérialisation des archives et métiers de l'archiviste*. Louvain-la-Neuve : Publications des archives de l'université catholique de Louvain.

BANA-BERGER, Françoise et HUC, Claude, 2011. Section 3 - *Les multiples visages du document numérique* [en ligne]. 22.11.2011. Support de cours : Cours « Module 7 - Gestion et archivage des documents numériques », Portail International Archivistique Francophone (PIAF) [Consulté le 18.08.2022]. Disponible à l'adresse : https://www.piaf-archives.org/sites/default/files/bulk_media/m07s03/co/section3_5.html

BÉCHARD, Lorène, FUENTES HASHIMOTO, Lourdes et VASSEUR, Édouard, 2020. *Les archives électroniques*. 2e édition. Paris : Association des archivistes français. Les petits guides des archives. ISBN 978-2-900175-10-1.

BELL, Mark et RANADE, Sonia, 2015. Traces through time: a case-study of applying statistical methods to refine algorithms for linking biographical data. *Proceedings of the First Conference on Biographical Data in a Digital World 2015*, Amsterdam, The Netherlands, April 9, 2015. N° 2015, pp. 9.

BELOVARI, Susanne, 2017. Expedited digital appraisal for regular archivists: an MPLP-type approach. *Journal of Archival Organization* [en ligne]. 3 avril 2017. Vol. 14, n° 1-2, pp. 55-77. [Consulté le 30 mars 2021]. DOI 10.1080/15332748.2018.1503014. Disponible à l'adresse : https://doi.org/10.1080/15332748.2018.1503014

BETA.GOUV, [s.d.]. Archifiltre. *beta.gouv* [en ligne]. [s.d.]. [Consulté le 7 janvier 2022]. Disponible à l'adresse : https://beta.gouv.fr/startups/archifiltre.html

BLEI, David M., 2012. Probabilistic topic models. *Communications of the ACM* [en ligne]. avril 2012. Vol. 55, n° 4, pp. 77-84. [Consulté le 10 décembre 2021]. DOI 10.1145/2133806.2133826. Disponible à l'adresse : https://dl.acm.org/doi/10.1145/2133806.2133826

BOLES, Frank et YOUNG, Julia, 1991. *Archival Appraisal*. New-York : Neal-Schuman.

BRETT, Megan R., 2012. Topic Modeling: A Basic Introduction. *Journal of Digital Humanities* [en ligne]. 2012. Vol. 2, n° 1. [Consulté le 10 décembre 2021]. Disponible à l'adresse : http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/

BUNN, Jenny, 2016. Archival description and automation: a brief history of going digital. *Archives and Records* [en ligne]. 2 janvier 2016. Vol. 37, n° 1, pp. 65-78. [Consulté le 24 mai 2021]. DOI 10.1080/23257962.2016.1145577. Disponible à l'adresse : https://doi.org/10.1080/23257962.2016.1145577

CENTRE INFORMATIQUE NATIONAL DE L'ENSEIGNEMENT SUPÉRIEUR (CINES), [s.d.]. Le modèle de référence : l'OAIS. *CINES* [en ligne]. [s.d.]. [Consulté le 5 janvier 2022]. Disponible à l'adresse : https://www.cines.fr/archivage/un-concept-des-problematiques/le-modele-de-reference-loais/

CHAUMARTIN, François-Régis et LEMBERGER, Pirmin, 2020. *Le traitement automatique des langues: comprendre les textes grâce à l'intelligence artificielle*. Malakoff : Dunod. InfoPro. ISBN 978-2-10-080188-6. 410.285

CLOUGH, Paul, TANG, Jiayu, HALL, Mark M. et WARNER, Amy, 2011. Linking archival data to location: a case study at the UK National Archives. WILLETT, Peter (éd.), *Aslib Proceedings* [en ligne]. 1 janvier 2011. Vol. 63, n° 2/3, pp. 127-147. [Consulté le 21 novembre 2021]. DOI 10.1108/00012531111135628. Disponible à l'adresse : https://doi.org/10.1108/00012531111135628

COMITÉ CONSULTATIF POUR LES SYSTÈMES DE DONNÉES SPATIALES (CCSDS), 2017. *Recommandation de pratiques pour les systèmes de données spatiales: Modèle de référence pour un Système ouvert d'archivage d'information (OAIS) Pratique Recommandée CCSDS 650.0-M-2 (F)* [en ligne]. Washington D.C. : CCSDS Secretariat. [Consulté le 27 novembre 2021]. Disponible à l'adresse : https://public.ccsds.org/Pubs/650x0m2%28F%29.pdf

CONFÉRENCE DES DIRECTRICES ET DIRECTEURS D'ARCHIVES SUISSES, 2021. *Schweizerische Archivstatistik 2013-2020 publiziert* [en ligne]. 2021. [Consulté le 29 novembre 2021]. Disponible à l'adresse : https://www.adk-cda.ch/fileadmin/user_upload/archivstatistik/Schweizerische_Archivstatistik_2013-2020_publiziert.xlsx

COOK, Terry, 1992. Mind Over Matter: Towards a New Theory for Archival Appraisal. In : CRAIG, Barbara L. et TAYLOR, Hugh A. (éd.), *The Archival Imagination: Essays in Honour of Hugh A. Taylor*. Ottawa : Association of Canadian Archivists. pp. 38-70. ISBN 1-895382-06-8.

COUTAZ, Gilbert, 2007. Archives publiques, archives privées: des solidarités nécessaires. *arbido* [en ligne]. 2007. [Consulté le 26 mai 2021]. Disponible à l'adresse : https://arbido.ch/fr/edition-article/2007/%C3%BCberlieferungsbildung-zusammenarbeit-und-gemeinsame-verantwortung-f%C3%BCr-transparenz/archives-publiques-archives-priv%C3%A9es-des-solidarit%C3%A9s-n%C3%A9cessaires

COUTAZ, Gilbert, 2014. La gestion des risques en termes de conservation de documents : du coffre-fort physique au coffre-fort numérique. *Dossier thématique* [en ligne]. 2014. pp. 36. [Consulté le 29 avril 2021]. Disponible à l'adresse : http://www.patrimoine.vd.ch/fileadmin/groups/19/PDF/Dossier_th%C3%A9matique_2014.pdf

COUTAZ, Gilbert, 2016. La croissance et la maîtrise des masses documentaires. *arbido* [en ligne]. 2016. N° 2016/3. [Consulté le 13 mai 2021]. 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-documentaireshttps://arbido.ch/fr/

COUTURE, Carol, 1996-1997. L'évaluation des archives. État de la question. *Archives*. [en ligne]. 1996-1997. Vol. 28, no. 1, pp. 3-21. [Consulté le 23 septembre 2021]. Disponible à l'adresse : https://www.archivistes.qc.ca/revuearchives/vol28_1/28-1-couture.pdf

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.

COUTURE, Carol et LAJEUNESSE, Marcel, 2014. *L'archivistique à l'ère du numérique: les éléments fondamentaux de la discipline*. Québec : Presses de l'Université du Québec. Collection Gestion de l'information. ISBN 978-2-7605-3998-3.

COUTURE, Carol et ROUSSEAU, Jean-Yves, 1982. *Les archives au XXe siècle. Une réponse aux besoins de l'administration et de la recherche*. Montréal : Université de Montréal.

Cryptogramme. *Grand dictionnaire terminologique*. [en ligne]. 2012. [Consulté le 14.01.2022]. Disponible à l'adresse : https://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=8353127

DESCHAMPS, Jacqueline, 2010. *Science de l'Information: de la discipline à l'enseignement.* Archives contemporaines. Paris. ISBN 978-2-8130-0028-6.

DOBRESKI, Brian, PARK, Jaihyun, LEATHERS, Alicia et QIN, Jian, 2020. Remodeling Archival Metadata Descriptions for Linked Archives. *International Conference on Dublin Core and Metadata Applications* [en ligne]. 30 mars 2020. pp. 1-11. [Consulté le 23 août 2021]. Disponible à l'adresse : https://dcpapers.dublincore.org/pubs/article/view/4223

Élimination. *Portail International Archivistique Francophone (PIAF)* [en ligne]. [s.d.]. [Consulté le 14.01.2022]. Disponible à l'adresse : https://www.piaf-archives.org/sites/default/files/bulk_media/glossaire/co/Module_glossaire_5.html#schId200

Empreinte numérique. *Grand dictionnaire terminologique* [en ligne]. 2012. [Consulté le 19 août 2022]. Disponible à l'adresse : https://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=8371028

FLORIDI, Luciano, 2010. Chapter 1: The information revolution. In : *Information: A Very Short Introduction*. Oxford University Press. Oxford. pp. 2-18. ISBN 978-0-19-955137-8.

Fonctionnalité. *Grand dictionnaire terminologique*. [en ligne] 2012. [Consulté le 14 janvier 2022]. Disponible à l'adresse : https://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=8386896

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.

FRANCEARCHIVES, 2021. Permaliens et identifiants pérennes. *FranceArchives*. [en ligne]. 31 août 2021. [Consulté le 13 janvier 2022]. Disponible à l'adresse: https://francearchives.fr/fr/article/339695117

GOLDMAN, Ben, 2017. Bridging the Gap: Taking Practical Steps Toward Managing Born-Digital Collections in Manuscript Repositories | Goldman | RBM: A Journal of Rare Books, Manuscripts, and Cultural Heritage. [en ligne]. 10 avril 2017. [Consulté le 23 octobre 2021]. DOI https://doi.org/10.5860/rbm.12.1.343. Disponible à l'adresse : https://rbm.acrl.org/index.php/rbm/article/view/343

GRACY, Karen F., 2015. Archival description and linked data: a preliminary study of opportunities and implementation challenges. *Archival Science* [en ligne]. 1 septembre 2015. Vol. 15, n° 3, pp. 239-294. [Consulté le 23 août 2021]. DOI 10.1007/s10502-014-9216-2. Disponible à l'adresse : https://doi.org/10.1007/s10502-014-9216-2

GREENE, Mark A., 2010. MPLP: It's Not Just for Processing Anymore. *The American Archivist* [en ligne]. 2010. Vol. 73, n° 1, pp. 175-203. [Consulté le 25 octobre 2021]. Disponible à l'adresse : https://www.jstor.org/stable/27802720

GREENE, Mark et MEISSNER, Dennis, 2005. More Product, Less Process: Revamping Traditional Archival Processing. *The American Archivist* [en ligne]. 1 septembre 2005. Vol. 68, n° 2, pp. 208-263. [Consulté le 13 mai 2021]. DOI 10.17723/aarc.68.2.c741823776k65863. Disponible à l'adresse : https://doi.org/10.17723/aarc.68.2.c741823776k65863

GUINCHAT, Claire et MENOU, Michel J., 1981. *Introduction générale aux sciences et techniques de l'information et de la documentation*. Paris : Presses de l'UNESCO. ISBN 978-92-3-201860-1. 00

GUYOT-JEANNING, Olivier, 1984. Tris et échantillonnages : empirisme et théorie. *Gazette des archives* [en ligne]. 1984. N° 124, pp. 5-26. [Consulté le 6 janvier 2021]. Disponible à l'adresse : https://www.persee.fr/doc/gazar_0016-5522_1984_num_124_1_2876

HARVEY, Ross et THOMPSON, Dave, 2010. Automating the appraisal of digital materials. *Library Hi Tech* [en ligne]. 1 janvier 2010. Vol. 28, n° 2, pp. 313-322. [Consulté le 25 mars 2021]. DOI 10.1108/07378831011047703. Disponible à l'adresse : https://doi.org/10.1108/07378831011047703

HIGGS, Emily, 2019. ml4arc – Machine Learning, Deep Learning, and Natural Language Processing Applications in Archives. *bloggERS!* [en ligne]. 4 septembre 2019. [Consulté le 23 novembre 2021]. Disponible à l'adresse : https://saaers.wordpress.com/2019/09/04/ml4arc-machine-learning-deep-learning-and-natural-language-processing-applications-in-archives/

HOOLAND, Seth van et COECKELBERGS, Mathias, 2018. Unsupervised Machine Learning for Archival Collections: Possibilities and Limits of Topic Modeling and Word Embedding. *Lligall - revista catalana d'Arxivística* [en ligne]. 2018. N° 41, pp. 73-90. [Consulté le 30 décembre 2021]. Disponible à l'adresse : https://arxivers.com/wp-content/uploads/2018/10/1.4_-Dossier_SVHooland_MCoeckelbergs.pdf

HUTCHINSON, Tim, 2020. Natural language Processing and Machine Learning as Practical Toolsets for Archival Processing. *Records Management Journal* [en ligne]. 16 mai 2020. Vol. 30, n° 2, pp. 155-174. [Consulté le 21 novembre 2021]. DOI 10.1108/RMJ-09-2019-0055. Disponible à l'adresse : https://www.emerald.com/insight/content/doi/10.1108/RMJ-09-2019-0055/full/html

KARENWARE, 2022. Karen's Directory Printer. *Karenware* [en ligne]. 2022. [Consulté le 8 janvier 2022]. Disponible à l'adresse : https://www.karenware.com/powertools/karens-directory-printer

KECSKEMÉTI, Charles et KÖRMENDY, Lajos, 2014. *Les écrits s'envolent la problématique de la conservation des archives papier et numériques*. Lausanne : Favre. ISBN 978-2-8289-1425-7.

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, n° 2, pp. 91-106. [Consulté le 30 mars 2021]. Disponible à l'adresse : https://www.jstor.org/stable/41102125

KIRSCHENBAUM, Matthew G., OVENDEN, Richard et REDWINE, Gabriela, 2010. pub149 : *Digital Forensics and Born-Digital Content in Cultural Heritage Collections* [en ligne]. Washington, D.C. : Council on Library and Information Resources. [Consulté le 7 janvier 2022]. Disponible à l'adresse : https://www.clir.org/pubs/reports/pub149/

LAMBERT, James et COTÉ, Louis, 1992. Les outils de travail en archivistique : la politique d'acquisition : pourquoi, comment, critères et exemples. *Archives*. 1992. Vol. 23, n° 3, pp. 3-50.

LEE, Christopher A., 2018. Computer-Assisted Appraisal and Selection of Archival Materials. In : *2018 IEEE International Conference on Big Data (Big Data)* [en ligne]. décembre 2018. pp. 2721-2724. [Consulté le 2 janvier 2022]. Disponible à l'adresse : https://ils.unc.edu/callee/p2721-lee.pdf

Linked Open Data. *W3C eGow Wiki* [en ligne]. Dernière modification de la page 7 juin 2010, à 18:53. [Consulté le 14 janvier 2022]. Disponible à l'adresse : https://www.w3.org/egov/wiki/Linked_Open_Data

Loi fédérale du 19 juin 1992 sur la protection des données (LPD ; 235.1). *Confédération suisse* [en ligne]. 19 juin 1992. Mise à jour le 1er mars 2019. [Consulté le 12 janvier 2022]. Disponible à l'adresse : https://www.fedlex.admin.ch/eli/cc/1993/1945_1945_1945/fr

Loi fédérale du 26 juin 1998 sur l'archivage (LAr ; 152.1). *Confédération suisse* [en ligne]. 26 juin 1998. Mise à jour le 1er mai 2013. [Consulté le 12 janvier 2022]. Disponible à l'adresse : https://www.fedlex.admin.ch/eli/cc/1999/354/fr

Loi sur l'archivage du 14 juin 2011 (LArch ; 432.11). *État de Vaud* [en ligne]. 14 juin 2011. [Consulté le 12 janvier 2022]. Disponible à l'adresse : https://prestations.vd.ch/pub/blv-publication/actes/consolide/432.11?key=1614767182310&id=f09bdf70-5553-4116-b2df-82226faac73a

Loi sur les archives publiques du 1er décembre 2000 (LArch ; rsGE B 2 15). *République et canton de Genève* [en ligne]. 1er décembre 2000. Mise à jour le 4 décembre 2018. [Consulté le 12 janvier 2022]. Disponible à l'adresse : https://ge.ch/archives/media/site_archives/files/imce/pdf/lois/20210412_larch.pdf

MAKHLOUF SHABOU, Basma, 2015. Fonction d'évaluation des archives : bilan sommaire des développements, des enjeux actuels et des défis futurs. In: COUTURE, Couture, LAJEUNESSE, Marcel et GAGNON-ARGUIN, Louise (éd.), *Panorama de l'archivistique contemporaine : évolution de la discipline et de la profession : mélanges offerts à Carol Couture*. Presses de l'Université du Québec. Québec. pp. 195-218.

MAKHLOUF SHABOU, Basma, 2021. *Introduction à l'OAIS* [document PDF]. 11 novembre 2021
Support de cours : Cours « Gouvernance des données : Introduction à l'OAIS », Haute école de gestion de Genève, filière sciences de l'information, année académique 2021-2022.

MAKHLOUF SHABOU, Basma, TIÈCHE, Julien, KNAFOU, Julien et GAUDINAT, Arnaud, 2020. Algorithmic methods to explore the automation of the appraisal of structured and unstructured digital data. *Records Management Journal* [en ligne]. 1 janvier 2020. Vol. 30, n° 2, pp. 175-200. [Consulté le 26 mai 2021]. DOI 10.1108/RMJ-09-2019-0049. Disponible à l'adresse : https://doi.org/10.1108/RMJ-09-2019-0049

MARCIANO, Richard, LEMIEUX, Victoria, HEDGES, Mark, ESTEVA, Maria, UNDERWOOD, William, KURTZ, Michael et CONRAD, Mark, 2018. Archival Records and Training in the Age of Big Data. In : PERCELL, Johnna, C. SARIN, Lindsay, T. JAEGER, Paul et CARLO BERTOT, John (éd.), *Re-envisioning the MLS: Perspectives on the Future of Library and Information Science Education* [en ligne]. Emerald Publishing Limited. pp. 179-199. Advances in Librarianship. [Consulté le 14 mai 2021]. ISBN 978-1-78754-884-8. Disponible à l'adresse : https://doi.org/10.1108/S0065-28302018000044B010

MATIENZO, Mark A., ROKE, Elizabeth Russey et CARLSON, Scott, 2017. Creating a Linked Data-Friendly Metadata Application Profile for Archival Description. *arXiv:1710.09688 [cs]* [en ligne]. 27 octobre 2017. [Consulté le 23 août 2021]. Disponible à l'adresse : http://arxiv.org/abs/1710.09688

Metadata [métadonnées]. *InterPARES Trust AI* [en ligne]. 2018. [Consulté le 14 janvier 2022]. Disponible à l'adresse : https://interparestrustai.org/terminology/term/metadata

MORISOD, Pascal, 2018. Des archives, des machines et des hommes, un heureux ménage…. *arbido* [en ligne]. 2018. [Consulté le 20 avril 2021]. Disponible à l'adresse : https://arbido.ch/fr/edition-article/2018/automatisierung-versprechen-oder-drohung/des-archives-des-machines-et-des-hommes-un-heureux-m%C3%A9nage-%C3%A0-trois

NATIONAL ARCHIVES AND RECORDS ADMINISTRATION, 2014. *Managing Government Records Directive - Automated Electronic Records Management Report/Plan* [en ligne]. National Archives and Records Administration. [Consulté le 26 décembre 2021]. Disponible à l'adresse : https://www.archives.gov/files/records-mgmt/prmd/A31report-9-19-14.pdf

NAUD, Dominique, 2019. Trois outils contribuant à l'archivage numérique. *Modernisation et archives* [en ligne]. 30 septembre 2019. [Consulté le 30 mars 2021]. Disponible à l'adresse : https://siaf.hypotheses.org/1033

Né numérique. *Portail International Archivistique Francophone (PIAF)* [en ligne]. 2011. [Consulté le 14.01.2022]. Disponible à l'adresse : http://www.piaf-archives.org/sites/default/files/bulk_media/m07s03/co/section3_5.html

OGUEY, Grégoire et SCHNEITER, Pascal, 2018. ArchiSelect, ou quand l'évaluation s'automatise. *arbido* [en ligne]. 2018. [Consulté le 23 mars 2021]. Disponible à l'adresse : https://arbido.ch/fr/edition-article/2018/automatisierung-versprechen-oder-drohung/archiselect-ou-quand-l%C3%A9valuation-sautomatise

Ontologie. *Grand dictionnaire terminologique*. [en ligne] 2012. [Consulté le 14 janvier 2022]. Disponible à l'adresse : https://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=8361952

Outil. *Grand dictionnaire terminologique*.[en ligne] 2012. [Consulté le 14 janvier 2022]. Disponible à l'adresse : https://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=26522836

PAYNE, Nathaniel, 2018. Stirring The Cauldron: Redefining Computational Archival Science (CAS) For The Big Data Domain. In : *2018 IEEE International Conference on Big Data (Big Data)*. décembre 2018. pp. 2743-2752.

PENN, Elaine, 2019. Appraising digital records, or Swimming in treacherous shoals. *The International Council on Archives Section on University and Research Institution Archives* [en ligne]. Dundee. 2 juillet 2019.. [Consulté le 23 avril 2022] Disponible en ligne : https://www.ica.org/sites/default/files/icasuv2019_penn.pdf

Personally identifiable information. InterPARES Trust AI [en ligne]. 2018. [Consulté le 14 janvier 2022]. Disponible à l'adresse : https://interparestrustai.org/terminology/term/personally%20identifiable%20information

PORTAIL INTERNATIONAL ARCHIVISTIQUE FRANCOPHONE, 2015. Glossaire. *Portail International Archivistique Francophone* [en ligne]. 31 octobre 2015.[Consulté le 27 mai 2021]. Disponible à l'adresse : https://www.piaf-archives.org/sites/default/files/bulk_media/glossaire/co/Module_glossaire.html

Préservation. *InterPARES Trust AI* [en ligne]. 2018. [Consulté le 14 janvier 2022]. Disponible à l'adresse : https://interparestrustai.org/terminology/term/preservation

Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère

personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données) (Texte présentant de l'intérêt pour l'EEE) (RGPD ; L 119/1). *Union européenne*. [en ligne]. 27 avril 2016. [Consulté le 12 janvier 2022]. Disponible à l'adresse : http://data.europa.eu/eli/reg/2016/679/oj/fra

ROLAN, Gregory, HUMPHRIES, Glen, JEFFREY, Lisa, SAMARAS, Evanthia, ANTSOUPOVA, Tatiana et STUART, Katharine, 2019. More human than human? Artificial intelligence in the archive. *Archives and Manuscripts* [en ligne]. 4 mai 2019. Vol. 47, n° 2, pp. 179-203. [Consulté le 24 mars 2021]. DOI 10.1080/01576895.2018.1502088. Disponible à l'adresse : https://doi.org/10.1080/01576895.2018.1502088

ROSS, Seamus, 2012. Digital Preservation, Archival Science and Methodological Foundations for Digital Libraries. *New Review of Information Networking* [en ligne]. mai 2012. Vol. 17, n° 1, pp. 43-68. [Consulté le 19 août 2021]. DOI 10.1080/13614576.2012.679446. Disponible à l'adresse : http://www.tandfonline.com/doi/abs/10.1080/13614576.2012.679446

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 7 janvier 2022]. ISBN 978-2-7605-1600-7. Disponible à l'adresse : https://hesge.scholarvox.com/catalog/book/docid/88801643

SAA Dictionary, [sans date]. [en ligne]. [Consulté le 13 janvier 2022]. Disponible à l'adresse : https://dictionary.archivists.org/index.html

SCHELLENBERG, Theodore R., 1964. *Modern Archives. Principles and Techniques*. Chicago : University of Chicago Press.

SCHELLENBERG, Theodore R., 1965. *Management of Archives*. New-York : Columbia University Press.

SHEIN, Cyndi, 2014. From Accession to Access: A Born-Digital Materials Case Study. *Journal of Western Archives* [en ligne]. 13 janvier 2014. Vol. 5, n° 1. DOI https://doi.org/10.26077/b3e2-d205. Disponible à l'adresse : https://digitalcommons.usu.edu/westernarchives/vol5/iss1/1

SIBILLE-DE GIMOÜARD, Claire et CAYA, Marcel, 2009. *Section5 – Description archivistique* [en ligne]. 28.09.2009. Support de cours : Cours « Module 6 – Traitement des archives définitives », Portail International Archivistique Francophone (PIAF) [Consulté le 14.01.2022]. Disponible à l'adresse : https://www.piaf-archives.org/sites/default/files/bulk_media/m06s5/co/06section5_web_1.html

SOCIALGOUV, 2022. Wiki Archifiltre. *GitHub* [en ligne]. 2022. [Consulté le 7 janvier 2022]. Disponible à l'adresse : https://github.com/SocialGouv/archifiltre

Tâche. *Grand dictionnaire terminologique*. [en ligne] 2012. [Consulté le 14 janvier 2022]. Disponible à l'adresse : https://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=8469938

THE NATIONAL ARCHIVES UK, 2016. *The application of technology- assisted review to born-digital records transfer, Inquiries and beyond* [en ligne]. [Consulté le 17 avril 2021]. Disponible à l'adresse : https://www.nationalarchives.gov.uk/documents/technology-assisted-review-to-born-digital-records-transfer.pdf

THE NATIONAL ARCHIVES UK, 2020a. *Market Research into AI/ML Tools for Document Selection and Classification* [en ligne]. First Phase Report. The National

Archives. [Consulté le 16 décembre 2021]. Disponible à l'adresse : https://cdn.nationalarchives.gov.uk/documents/phase-1-market-research-ai.pdf

THE NATIONAL ARCHIVES UK, 2020b. *DROID : User Guide* [en ligne]. mai 2020. [Consulté le 9 janvier 2022]. Disponible à l'adresse : https://cdn.nationalarchives.gov.uk/documents/information-management/droid-user-guide.pdf

THE NATIONAL ARCHIVES UK, 2021a. Download DROID: file format identification tool. *The National Archives Official Homepage* [en ligne]. 2021. [Consulté le 7 janvier 2022]. Disponible à l'adresse : https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/

THE NATIONAL ARCHIVES UK, 2021b. *Using AI for Digital Records Selection in Government : Guidance for records managers based on an evaluation of current marketplace solutions* [en ligne]. [Consulté le 18 novembre 2021]. Disponible à l'adresse : https://cdn.nationalarchives.gov.uk/documents/using-ai-digital-selection-in-government.pdf

TRACE, Ciaran B., 2021. Archival infrastructure and the information backlog. *Archival Science* [en ligne]. 21 juillet 2021. [Consulté le 29 novembre 2021]. DOI 10.1007/s10502-021-09368-x. Disponible à l'adresse : https://doi.org/10.1007/s10502-021-09368-x

VAN DER MAREN, Jean-Marie, 1996. *Méthodes de recherche pour l'éducation* [en ligne]. Presses de l'Université de Montréal et de Boeck. [Consulté le 4 janvier 2022]. ISBN 978-2-8041-2364-2. Disponible à l'adresse : https://papyrus.bib.umontreal.ca/xmlui/handle/1866/4688

VELLINO, André et ALBERTS, Inge, 2016. Assisting the appraisal of e-mail records with automatic classification. *Records Management Journal* [en ligne]. 1 janvier 2016. Vol. 26, n° 3, pp. 293-313. [Consulté le 30 mars 2021]. DOI 10.1108/RMJ-02-2016-0006. Disponible à l'adresse : https://doi.org/10.1108/RMJ-02-2016-0006

WEILL, Georges, 1990. Les mutations de l'archivistique contemporaine. *Gazette des archives* [en ligne]. 1990. Vol. 149, n° 1, pp. 107-118. [Consulté le 2 décembre 2021]. DOI 10.3406/gazar.1990.3151. Disponible à l'adresse : https://www.persee.fr/doc/gazar_0016-5522_1990_num_149_1_3151

YOUNG, Tom, HAZARIKA, Devamanyu, PORIA, Soujanya et CAMBRIA, Erik, 2018. Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. *IEEE Computational Intelligence Magazine*. août 2018. Vol. 13, n° 3, pp. 55-75.

Annexe 1 : Tableau synoptique des outils

Nom	Type	Accessibilité	Développeur	Année début	Année fin	Fonctionnalités	Description	Référence(s) bibliographique(s)	Liens
Access to Memory (AtoM)	Outil	Libre	Artefactual Systems	2007		Modèle de description standardisé; Contrôle d'autorité; Contrôle des droits d'accès; Contrôle du vocabulaire; Création de copies d'accès; Recherche générale	"AtoM (short for Access to Memory) is a web-based, open source application for standards-based archival description and access. The application is multilingual and multi-repository. First commissioned by the International Council on Archives (ICA) to make it easier for archival institutions worldwide to put their holdings online using the ICA's descriptive standards, the project has since grown into an internationally used community-driven project." (github AtoM)		https://www.accesstomemory.org/ ; https://github.com/artefactual/atom (consultés le 23.12.2021)
Adlib Elevate	Outil	Propriétaire	Adlib			Dédoublonnage; Reconnaissance optique de caractères imprimés; Classification et catégorisation de documents (NLP); Extraction de métadonnées; Recherche plein-texte	"Elevate is a cloud-based off-the-shelf record management product developed by Adlib. Although the platform is designed for ease of use by Records Managers, it also provides functionality for data scientists to engage with the model building process. It can be used to transform large collections of unstructured data into structured data to carry out business processes » (The National Archives UK 2021)	The National Archives UK 2021	https://cdn.nationalarchives.gov.uk/documents/deloitte-adlib-national-archives.pdf ; https://www.adlibsoftware.com/platform (consultés le 23.12.2021)
Amazon Textract	Outil	Propriétaire	Amazon Web Services			Reconnaissance optique de caractères imprimés; Reconnaissance optique de caractères manuscrits	"Amazon Textract is a machine learning (ML) service that automatically extracts text, handwriting, and data from scanned documents. It goes beyond simple optical character recognition (OCR) to identify, understand, and extract data from forms and tables. Today, many companies manually extract data from scanned documents such as PDFs, images, tables, and forms, or through simple OCR software that requires manual configuration (which often must be updated when the form changes). To overcome these manual and expensive processes, Textract uses ML to read and process any type of document, accurately extracting text, handwriting, tables, and other data with no manual effort. You can quickly automate document processing and act on the information extracted, whether you're automating loans processing or extracting information from invoices and receipts. Textract can extract the data in minutes instead of hours or days. Additionally,	The National Archives UK 2021	https://aws.amazon.com/textract/ (consulté le 20.12.2021)

Nom	Type	Accessibilité	Développeur	Année début	Année fin	Fonctionnalités	Description	Référence(s) bibliographique(s)	Liens
							you can add human reviews with Amazon Augmented AI to provide oversight of your models and check sensitive data." (site Amazon Textract)		
Ant Renamer	Outil	Libre	Antoine Potten (particulier)	2000	2015	Renommage de fichiers; Changement d'extension	"Ant Renamer est un programme libre et gratuit (vraiment libre, càd que le code source est disponible) permettant de renommer facilement de grandes quantités de fichiers et dossiers selon des critères définis. Il supporte les noms en Unicode." (site Ant Renamer)	coline1 2019	https://www.antp.be/software/renamer/fr ; https://github.com/carry0987/Ant-Renamer (consultés le 27.12.2021)
ArchExtract	Outil	Libre	Bancroft Library (University of California Berkeley)	2015		Reconnaissance d'entités nommées (NLP); Topic modeling (NLP); Reconnaissance d'entités nommées (NLP)	"ArchExtract is web application that enables archivists and researchers to perform topic modeling, keyword and named entity extraction on a text collection. The web application automates and packages a number of existing natural language processes and algorithms for the researcher or archivist. Using automated text analysis as the starting point, ArchExtract illuminates the scope and content of a digital text collection and provides an web-based interface for text exploration." (github ArchExtract)	Elings 2016; Elings 2017; Hutchinson 2020	https://github.com/j9recurses/archextract (consulté le 27.12.2021)
ArchiClass	Outil	Propriétaire	AEN (Projet AENEAS)	2015		Création d'un calendrier de conservation; Création d'un plan de classement; Importation d'un plan de classement préexistant; Ajout de métadonnées descriptives techniques	"C'est un logiciel d'élaboration de plans d'archivage. Il permet de créer un cadre de classement et d'y adjoindre différentes métadonnées archivistiques, en particulier les durées d'utilité et le sort final pressenti." (site AEN)	Oguey et Schneiter 2018;	https://archiclass.ch/ ; https://www.ne.ch/autorites/DESC/SCNE/archives-etat/numerique/Pages/76-BoiteOutils.aspx (consultés le 27.12.2021)

Nom	Type	Accessibilité	Développeur	Année début	Année fin	Fonctionnalités	Description	Référence(s) bibliographique(s)	Liens
Archifiltre	Outil	libre	Fabrique des ministères sociaux (France)	2018		Visualisation de l'arborescence; Ajout de métadonnées descriptives intellectuelles; Dédoublonnage; Création d'un rapport synthétique; Création d'un rapport de récolement; Remaniement de l'arborescence; Exportation vers un SAE; Création d'un bordereau d'élimination; Garantie de la valeur probante (empreinte digitale numérique)	"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)." (github Archifiltre)	Makhlouf et al. 2020; Naud 2019	https://github.com/SocialGouv/archifiltre/wiki/Wiki-Archifiltre (consulté le 27.12.2021)
Archivematica	Outil	Libre	artefactual	2009		Renommage de fichiers	"Archivematica is a digital preservation platform that ingests content, performs configurable preservation actions and generates standards-based, self-documenting Archival Information Packages for long-term storage. Archivematica automates standard digital preservation activities such as ingestion, checksum generation, format identification, format validation, metadata extraction, format conversion and placement in archival storage. Content can be re-ingested and new workflows initiated to accommodate new format migrations, metadata updates or other preservation actions. The user can interact with the system via a web-based dashboard, but configuration options can be used to fully automate all aspects of the workflow." (site Archivematica, "Information summary") "Archivematica is an integrated suite of open-source software tools that allows users to process digital objects from ingest to access in compliance with the ISO-OAIS functional model. Users monitor and control ingest and preservation micro-services via a web-based dashboard. Archivematica uses METS, PREMIS, Dublin Core, the Library of Congress BagIt specification and other recognized standards to generate trustworthy, authentic, reliable and system-	Hutchinson 2020; Shein 2014	https://www.archivematica.org/fr/ ; https://www.artefactual.com/wp-content/uploads/2019/07/Archivematica-information-summary-2019.pdf (consultés le 23.12.2021)

Nom	Type	Accessibilité	Développeur	Année début	Année fin	Fonctionnalités	Description	Référence(s) bibliographique(s)	Liens
							independent Archival Information Packages (AIPs) for storage in your preferred repository." (site Archivematica)		
ArchivesSpace	Outil	Libre	Lyrasis	2013		Remaniement de l'arborescence; Contrôle des droits d'accès; Contrôle d'autorité; Journalisation; Ajout de métadonnées descriptives intellectuelles; Ajout de métadonnées techniques	"ArchivesSpace is an open source, web application for managing archives information. The application is designed to support core functions in archives administration such as accessioning; description and arrangement of processed materials including analog, hybrid, and born-digital content; management of authorities (agents and subjects) and rights; and reference service. The application supports collection management through collection management records, tracking of events, and a growing number of administrative reports. The application also functions as a metadata authoring tool, enabling the generation of EAD, MARCXML, MODS, Dublin Core, and METS formatted data." (site Archivesspace)	Shein 2014	https://archivesspace.org/ (consulté le 23.12.2021)
AWS Comprehend	Outil	Propriétaire	Amazon; Kainos			Extraction de texte; Dédoublonnage; Reconnaissance d'entités nommées (NLP); Topic modeling (NLP); Analyse des sentiments (NLP)	"Amazon Comprehend est un service de traitement du langage naturel (NLP) qui utilise le machine learning (ML) pour découvrir des informations et des relations utiles dans un texte." (site AWS Comprehend)	The National Archives UK 2021	https://aws.amazon.com/fr/comprehend/ (consulté le 27.12.2021)

Nom	Type	Accessibilité	Développeur	Année début	Année fin	Fonctionnalités	Description	Référence(s) bibliographique(s)	Liens
Azure Cognitive Search	Outil	Propriétaire	Microsoft			Sensitivity Review (NLP); Reconnaissance d'images; Reconnaissance d'entités nommées (NLP); Reconnaissance optique de caractères imprimés;	"Azure Cognitive Search is the only cloud search service with built-in AI capabilities that enrich all types of information to help you identify and explore relevant content at scale. Use cognitive skills for vision, language, and speech, or use custom machine learning models to uncover insights from all types of content. Azure Cognitive Search also offers semantic search capability, which uses advanced machine learning techniques to understand user intent and contextually rank the most relevant search results. Spend more time innovating and less time maintaining a complex cloud search solution." (site Azure microsoft)	The National Archives UK 2021	https://cdn.nationalarchives.gov.uk/documents/adatis-azure-national-archives.pdf ; https://azure.microsoft.com/en-us/services/search/ (consulté le 23.12.2021)
Better File Rename	Outil	Propriétaire	Publicspace.net	2003		Renommage de fichiers	"Better File Rename's huge array of renaming options is organized into 10 intuitive categories that cover all the text, character, position, conversion and truncation features that you would expect from a great file renamer. On top of this, Better File Rename provides advanced features that answer the prayers of many professionals and hobbyists alike. [...] Today's media files come with an abundance of additional information that cannot be glanced from the often meaningless file names themselves. Better File Rename allows you to leverage this meta-data to create more meaningful file names using its tag-based renaming feature. Our renaming engine can read an extensive array of photo, image, music, movie, camera, lens and location meta-data and you can combine this information to implement any naming scheme you can imagine." (site Public Space, Better File Rename)	Shein 2014, p. 20	https://www.publicspace.net/windows/BetterFileRename/index.html (consulté le 23.12.2021)

Nom	Type	Accessibilité	Développeur	Année début	Année fin	Fonctionnalités	Description	Référence(s) bibliographique(s)	Liens
Beyond Compare	Outil	Propriétaire	Scooter software			Comparaison de fichiers texte	"You can compare entire drives and folders at high speed, checking just sizes and modified times. Or, thoroughly verify every file with byte-by-byte comparisons. FTP sites, cloud storage, and zip files are integrated seamlessly, and powerful filters allow you to limit what you see to only what you're interested in. Once you've found specific files you're interested in, Beyond Compare can intelligently pick the best way to compare and display them. Text files can be viewed and edited with syntax highlighting and comparison rules tweaked specifically for documents, source code, and HTML. The text contents of Microsoft Word .doc and Adobe .pdf files can also be compared but not edited. Data files, executables, binary data, and images all have dedicated viewers as well, so you always have a clear view of the changes. Beyond Compare's merge view allows you to combine changes from two versions of a file or folder into a single output. Its intelligent approach allows you to quickly accept most changes while carefully examining conflicts. Color coding and section highlighting allow you to accept, reject, or combine changes, simply and easily. When merging files you can change any line in the output with the built-in syntax-highlighting editor. By using Beyond Compare's powerful file type support and ability to favor changes from one file, you can trivially accept many changes without even seeing them. You can use Beyond Compare directly from most version control systems, giving you all of the powerful comparing and merging support you need when you need it most. Integrated source control commands are also available, allowing you to check in and check out files without interrupting your work." (site Scootersoftware)	Shein 2014, p. 19	https://www.scootersoftware.com/; https://www.scootersoftware.com/features.php?zz=features_list (consultés le 23.12.2021)
BitCurator NLP	Outil	Libre	School of Information and Library Science at the University of North Carolina at Chapel Hill (SILS); Maryland Institute for Technology in the Humanities (MITH);	2016		Extraction de texte; Topic modeling (NLP); Reconnaissance d'entités nommées (NLP); Comparaison de fichiers texte; Prévisualisation de fichiers	"BitCurator NLP project personnel developed software for collecting institutions to extract, analyze, and produce reports on features of interest in text extracted from born-digital materials contained in collections. The software uses existing natural language processing software libraries to identify and report on those items likely to be relevant to ongoing preservation, information organization, and access activities. These may include entities (e.g. persons, places, and organizations), potential relationships among entities (for example, by describing those entities that appear together within documents or set of documents), and topic models to provide insight into how	Goodman 2019; Hutchinson 2020; Lee 2018; Shein 2014	https://bitcurator.net/bitcurator-nlp/ ; https://github.com/BitCurator/bitcurator-nlp/ (consultés le 23.12.2021)

Nom	Type	Accessibilité	Développeur	Année début	Année fin	Fonctionnalités	Description	Référence(s) bibliographique(s)	Liens
			BitCurator Consortium				concepts are naturally clustered within the documents." (github BitCurator NLP)		
Bulk Reviewer	Outil	Libre		2019		Sensitivity Review (NLP); Annotation de document; Création d'un rapport de recherche; Sensitivity Review (NLP); Topic modeling (NLP)	"Bulk Reviewer is an Electron desktop application that aids in identification, review, and removal of sensitive files in directories and disk images. Bulk Reviewer scans directories and disk images for personally identifiable information (PII) and other sensitive information using bulk_extractor, a best-in-class digital forensics tool." (github Bulk Reviewer)	Hutchinson 2020	https://github.com/bulk-reviewer/bulk-reviewer (consulté le 21.12.2021)
bulk_extractor	Outil	Libre		2013		Extraction de texte; Sensitivity Review (NLP)	"bulk_extractor is a high-performance digital forensics exploitation tool. It is a "get evidence" button that rapidly scans any kind of input (disk images, files, directories of files, etc) and extracts structured information such as email addresses, credit card numbers, JPEGs and JSON snippets without parsing the file system or file system structures. The results are stored in text files that are easily inspected, searched, or used as inputs for other forensic processing. bulk_extractor also creates histograms of certain kinds of features that it finds, such as Google search terms and email addresses, as previous research has shown that such histograms are especially useful in investigative and law enforcement applications." (github bulk_extractor)	Hutchinson 2020	https://github.com/simsong/bulk_extractor (consulté le 21.12.2021)
CollectiveAccess	Outil	Libre	Whirl-i-Gig	2012		Création d'un rapport de recherche; Création d'un rapport de récolement; Recherche générale; Visualisation de chronologie; Visualisation par cartes géographiques	"CollectiveAccess is open-source collections management and presentation software designed for museums, archives, and special collections also increasingly used by libraries, corporations and non-profits. It is designed to handle large, heterogeneous collections that have complex cataloguing requirements and require support for a variety of metadata standards and media formats. CollectiveAccess is a collaboration between Whirl-i-Gig and partner institutions in North America and Europe with projects in 5 continents. The software is freely available under the open source GNU General Public License, meaning it's not only free to download and use but that users are encouraged to share and distribute code." (site CollectiveAccess)		https://collectiveaccess.org/ ; https://github.com/collectiveaccess (consultés le 20.12.2021)

Nom	Type	Accessibilité	Développeur	Année début	Année fin	Fonctionnalités	Description	Référence(s) bibliographique(s)	Liens
DiAGRAM	Outil	Libre	The National Archives UK	2020		Evaluation des risques archivistiques	"DiAGRAM is an online tool designed to help archivists manage the risks to their digital collections. By answering a set of questions relating to archives such as storage media, system security and technical skills, the tool will use statistical methods to calculate the probability that your digital material is preserved." (site DiAGRAM)		https://nationalarchives.shinyapps.io/DiAGRAM/ (consulté le 18.12.2021)
Digital Record Object Identification (DROID)	Outil	Libre	The National Archives UK	2005		Identification de format; Création d'un rapport de récolement	"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 National Archives UK)	Lee 2018	https://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/ (consulté le 18.12.2021)
Docuteam Packer	Outil	Libre	Docuteam	2013		Ajout de métadonnées descriptives techniques; Ajout de métadonnées descriptives intellectuelles	"docuteam packer est une application Java permettant de préparer des données pour une archive numérique. docuteam packer transforme vos fichiers en un paquet d'informations de soumission (SIP). Il contient les données de la soumission ainsi que des métadonnées techniques, structurelles et descriptives." (site docuteam)		https://docs.docuteam.ch/packer/6.0/fr/index (consulté le 20.12.2021)
Duke DataAccessioner	Outil	Libre	David M. Rubenstein Rare Book & Manuscript Library	2008	2017	Migration de support; Dédoublonnage; Extraction de métadonnées; Compilation de métadonnées (en format XML)	"The Data Accessioner (DA) is a simple tool, with an easy-to-use graphic interface, for migrating content between media while also: creating and validating checksums; gathering metadata (via FITS); and compiling an XML metadata file (with the option to include Dublin Core metadata as of v 1.0) for future reference. The DataAccessioner was built out of the need for a simple GUI interface to allow the Duke University Rare Book, Manuscripts, & Special Collections Library (now the David M. Rubenstein Rare Book & Manuscript Library) Technical Services staff an easy way of migrating data off disks and onto a file server for basic preservation, further appraisal, arrangement, & description. It also provides a way to integrate common metadata tools at the time of migration rather than after the fact. With a simplified interface and being written in Java it is intended to be easily adopted by smaller institutions with little or no IT staff support." (site dataaccessioner)	Shein 2014, p. 5; Goldman 2011, p. 21	http://dataaccessioner.org/ (consulté le 21.12.2021)

Nom	Type	Accessibilité	Développeur	Année début	Année fin	Fonctionnalités	Description	Référence(s) bibliographique(s)	Liens
Elsasticsearch	Outil	Libre	Elastic	2012		Recherche plein texte; Recherche à facettes	"Elasticsearch is the distributed, RESTful search and analytics engine […]. You can use Elasticsearch to store, search, and manage data" (github elasticsearch)		https://github.com/elastic/elasticsearch ; https://www.elastic.co/fr/app-search/ (consultés le 31.12.2021)
ePADD	Outil	Libre	Stanford University Libraries	2015		Extraction d'informations contextuels (NLP); Dédoublonnage; Reconnaissance d'entités nommées (NLP); Sensitivity Review (NLP); Topic modeling (NLP); Annotation de document	"ePADD is a software package developed by Stanford University's Special Collections & University Archives that supports archival processes around the appraisal, ingest, processing, discovery, and delivery of email archives." (github epadd)	Hutchinson 2020; Lee 2018; Schneider et al. 2019	https://github.com/ePADD/epadd ; https://library.stanford.edu/projects/epadd (consultés le 27.12.2021)
EpubCheck	Outil	Libre	Daisy Consortium	2013		Validation de format	"EPUBCheck is a tool to validate the conformance of EPUB publications against the EPUB specifications." (github epubcheck)	Conraux 2020	https://github.com/w3c/epubcheck (consulté le 20.12.2021)
everteam.archive	Outil	Propriétaire	Kyocera Document Solutions			Garantie de la valeur probante (empreinte digitale numérique); Conversion de format; Journalisation	"In a context of exploding digital content volumes, the implementation of retention schedules and policies guaranteeing thorough information governance is essential. Through a single solution, manage the archiving of your information assets, regardless of their initial support." (site everteam)		https://www.everteam.com/en/electronic-archiving-solutions/ (consulté le 22.12.2021)
everteam.discover	Outil	Propriétaire	Kyocera Document Solutions			Recherche plein-texte; Reconnaissance d'entités nommées (NLP); Sensitivity Review (NLP);	"The everteam.discover solution is a platform for analyzing and enriching your content in order to simplify your regulatory compliance, to preserve your sensitive data, to migrate your documents or simply to cleanse and purify your information capital." (site everteam)		https://www.everteam.com/en/data-analytics-solutions/ (consulté le 22.12.2021)
File Information Tool Set (FITS)	Outil	Libre	Harvard Library			Extraction de métadonnées; Identification de format	"The File Information Tool Set (FITS) identifies, validates and extracts technical metadata for a wide range of file formats. It acts as a wrapper, invoking and managing the output from several other open source tools. Output from these tools are converted into a	Harvey et Thompson 2010	https://projects.iq.harvard.edu/fits/home (consulté le 21.12.2021)

Nom	Type	Accessibilité	Développeur	Année début	Année fin	Fonctionnalités	Description	Référence(s) bibliographique(s)	Liens
							common format, compared to one another and consolidated into a single XML output file. FITS is written in Java and is compatible with Java 1.8 or higher." (site projects harvard)		
Flora	Outil	Propriétaire	Decalog			Contrôle des droits d'accès; Attribution d'identifiants pérennes; Importation de documents; Recherche à facettes; Application des délais de conservation; Journalisation	Depuis plus de 30 ans, Decalog Flora propose à des institutions de toutes natures, tailles et organisations de gérer de manière unifiée leurs ressources documentaires : gestion des collections patrimoniales (musées, fondations, collections privées), documentation (texte, images, vidéos, sons), bibliographie UNIMARC conforme aux évolutions de la transition bibliographique et données archivistiques au formats normés ISAD(G), ISAAR(CPF) et XML-EAD. (site Flora)		https://flora.decalog.net/ (consulté le 30.12.2021)
Forensic Recovery of Evidence Device (FRED)	Outil	Propriétaire	Digital Intelligence			Bloquage d'écriture	"FRED systems set the standard for forensic acquisition and analysis workstations. The quality, features, performance, and overall capability are second to none. Buying a FRED system means making an investment in your ability to solve every investigation. [...] FRED systems are designed and built from the ground up as high performance, forensic acquisition, analysis and processing platforms" (site Digital Intelligence)	Shein 2014, p. 3; Krichenbaum 2010, p. 21, 30, 83	https://digitalintelligence.com/products/fred (consulté le 22.12.2021)
Forensic ToolKit (FTK)	Outil	Propriétaire	AccessData (Exterro)			Décryptage de fichiers; Craquer mot de passe; Reconnaissance d'images; Visualisation de l'arborescence	"Create full-disk forensic images and process a wide range of data types from many sources, from hard drive data to mobile devices, network data and Internet storage, all in a centralized, secure database. FTK® processes and indexes data upfront, eliminating wasted time waiting for searches to execute. Cut down on OCR time by up to 30% with our efficient OCR engine. Decrypt files, crack passwords, and build reports with a single solution. Recover passwords from over 100+ applications. Decrypt a computer drive encrypted by the latest version of McAfee Drive Encryption and features L01 export support, which eases the workflow of users when data must be used within multiple tools. Parse registry files and Windows system information files in an easy to read, interactive and reportable tab. Label, bookmark and export individual objects per category, allowing for easy searching, filtering and reporting. Locate, manage and filter mobile data more easily with a dedicated	Shein 2014; Kirschenbaum 2010, p. 18	https://www.exterro.com/forensic-toolkit (consulté le 21.12.2021)

Nom	Type	Accessibilité	Développeur	Année début	Année fin	Fonctionnalités	Description	Référence(s) bibliographique(s)	Liens
							mobile tab. Use the message application filter to quickly isolate data from message applications like WhatsApp or Facebook. Collect, process and analyze datasets containing Apple file systems that are encrypted, compressed or deleted. FTK® Supports decryption of File Vault 2 from the APFS file system, as well as importing and parsing of AFF4 images created from Mac® computers (generated by third-party solutions like MacQuisition by BlackBag). Visualization technology that helps you get a clearer picture of events by displaying your data in timelines, cluster graphs, pie charts, geolocations, and more. Dig deeper and view all EXIF data, including location, make and model of the device used to capture images or video" (site exterro)		
Format Identification for Digital Objects (FIDO)	Outil	Libre	Open Preservation Foundation	2010		Identification de format	"Format Identification for Digital Objects (FIDO) is a Python command-line tool to identify the file formats of digital objects. It is designed for simple integration into automated work-flows." (github fido)	Hutchinson 2020	https://fido.openpreservation.org/ ; https://github.com/openpreserve/fido
InSight	Outil	Propriétaire	Iron Mountain			Importation de documents; Ajout de métadonnées descriptives intellectuelles; Ajout de métadonnées techniques; Recherche générale	"Iron Mountain InSight provides the industry's most advanced Content Services Platform (CSP) to help organizations unlock the hidden value of their content —wherever it resides — through modern, AI-infused content-driven applications, and federated search capabilities." (site Iron Mountain)	The National Archives UK 2021	https://www.ironmountain.com/services/content-service-platform (consulté le 23.12.2021)
Iso Buster	Outil	Propriétaire	Smart Projects			Récupération de données; Changement d'extension	"IsoBuster est un outil de récupération de médias hautement spécialisé et simple d'utilisation. Il prend en charge tous les types de médias et tous leurs système de fichiers communs. Démarrez IsoBuster, insérez un disque (dans le cas d'un disque optique), sélectionnez le lecteur (s'il ne l'est pas déjà) et laissez IsoBuster charger le média. IsoBuster vous montrera immédiatement toutes les pistes et sessions présentes sur le média, combinés à tous les systèmes de fichiers présents. De cette façon, vous avez facilement accès, comme l'Explorateur, à tous les fichiers et répertoires par système de fichier. Au lieu d'être limité au système de fichiers que le système choisit pour vous, vous avez accès à "l'image complète". Accédez aux données des sessions précédentes, aux données que	Shein 2014	https://www.isobuster.com/fr/isobuster.php (consulté le 21.12.2021)

Nom	Type	Accessibilité	Développeur	Année début	Année fin	Fonctionnalités	Description	Référence(s) bibliographique(s)	Liens
							votre système (par ex. Windows) ne voit pas ou vous cache, etc." (site Isobuster)		
JHOVE	Outil	Libre	JSTOR; Harvard university library; Open Preservation Foundation	2015		Identification de format; Validation de format	"JHOVE (the JSTOR/Harvard Object Validation Environment, pronounced "jove") is an extensible software framework for performing format identification, validation, and characterization of digital objects." (github JHOVE)	Conraux 2020	https://github.com/openpreserve/jhove/ ; https://jhove.openpreservation.org/ (consultés le 20.12.2021)
Jpylyzer	Outil	Libre	Open Preservation Foundation	2016		Validation de format; Extraction de métadonnées	"P2 (JPEG 2000 Part 1) validator and properties extractor. Jpylyzer was specifically created to check that a JP2 file really conforms to the format's specifications. Additionally jpylyzer is able to extract technical characteristics." (site jpylyzer)	Conraux 2020	https://jpylyzer.openpreservation.org/ ; https://github.com/openpreserve/jpylyzer (consultés le 21.12.2021)
JWAT-Tools	Outil	Libre	Danish Royal Library	2014		Identification de format	"JWAT-Tools is an extension to the JWAT [Java Web Archive Toolkit] utility libraries." (github JWAT-tools)	Conraux 2020	https://github.com/netarchivesuite/jwat-tools (consulté le 21.12.2021)
Karen's Directory Printer	Outil	Libre	Karen's Power Tools (Karen Kenworthy & Joe Winett)	1997		Création d'un rapport de récolement	"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 Karenware)	Shein 2014, p. 18	https://www.karenware.com/powertools/karens-directory-printer (consulté le 22.12.2021)
log2timeline plaso	Outil	Libre	Apache Software Foundation	2016		Extraction de métadonnées; Visualisation de chronologie	"Plaso (Plaso Langar Að Safna Öllu), or super timeline all the things, is a Python-based engine used by several tools for automatic creation of timelines." (github plaso)	Lee 2018	https://github.com/log2timeline/plaso/ (consulté le 21.12.2021)

Nom	Type	Accessibilité	Développeur	Année début	Année fin	Fonctionnalités	Description	Référence(s) bibliographique(s)	Liens
Luminoso Daylight	Outil	Propriétaire	Luminoso	2010		Analyse des sentiments (NLP)	"Luminoso Daylight is a text analysis application for conversational use, such as support tickets, open-ended survey responses, and product reviews." (site Luminoso)	Hutchinson 2020	https://www.luminoso.com/dalight (consulté le 21.12.2021)
MediaInfo	Outil	Libre	MediaArea			Extraction de métadonnées	"MediaInfo fournit des informations techniques et les tags à propos de vos fichiers video et audio." (site Mediainfo)	Conraux 2020	https://mediaarea.net/fr/MediaInfo ; https://github.com/MediaArea/MediaInfo (consultés le 22.12.2021)
Memory Using Email (MUSE)	Outil	Libre	MobiSocial Computing Laboratory (Stanford University)			Conversion de format; Analyse des sentiments; Recherche à facettes; Reconnaissance d'entités nommées (NLP); Visualisation de chronologie; Topic modeling (NLP)	"Muse is a research tool from Stanford Computer Science for browsing large email archives. It was originally meant for people to browse their own long-term email archives. We have now started adapting it for journalists, archivists and researchers." (site MUSE)	Makhlouf Shabou 2015	https://mobisocial.stanford.edu/muse/ (consulté le 20.12.2021)
Metadata Extraction Tool	Outil	Libre	National Library of New Zealand	2003		Extraction de métadonnées; Ajout de métadonnées descriptives techniques; Compilation de métadonnées	"The Metadata Extraction Tool was developed by the National Library of New Zealand to programmatically extract preservation metadata from a range of file formats like PDF documents, image files, sound files Microsoft office documents, and many others. [...] The Tool builds on the Library's work on digital preservation, and its logical preservation metadata schema. It is designed to: automatically extracts preservation-related metadata from digital files ; output that metadata in a standard format (XML) for use in preservation activities. The Tool was designed for preservation processes and activities, but can be used to for other tasks, such as the extraction of metadata for resource discovery. " (site Metadata Extractor Tool)	Kim 2006, p. 98; Kirschenbaum 2010, p. 18	http://meta-extractor.sourceforge.net/ (consulté le 21.12.2021)
Natural Language Toolkit (NLTK)	Outil	Libre	University of Pennsylvania	2001		Topic modeling (NLP); Reconnaissance d'entités nommées (NLP)	"NLTK -- the Natural Language Toolkit -- is a suite of open source Python modules, data sets, and tutorials supporting research and development in Natural Language Processing." (github NTLK)	Lee 2018	https://github.com/nltk/nltk ; https://www.nltk.org/ (consultés le 21.12.2021)

Nom	Type	Accessibilité	Développeur	Année début	Année fin	Fonctionnalités	Description	Référence(s) bibliographique(s)	Liens
Open Calais	Outil	Libre	Thomson Reuters	2008		Extraction d'informations contextuelles (NLP); Sensitivity Review (NLP); Topic modeling (NLP); Reconnaissance d'entités nommées (NLP); Ajout de métadonnées descriptives intellectuelles	"Intelligent Tagging uses natural language processing, text analytics and data-mining technologies to derive meaning from vast amounts of unstructured content. It's the fastest, easiest and most accurate way to tag the people, places, facts and events in your data, and then assign financial topics and themes to increase your content's value, accessibility and interoperability." (site Open Calais)	Gracy 2015, p. 261	https://www.refinitiv.com/en/products/intelligent-tagging-text-analytics (consulté le 20.12.2021)
OpenNLP	Outil	Libre	Apache Software Foundation	2004		Reconnaissance d'entités nommées (NLP); Identification de langue	"The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text." (github openNLP)	Lee 2018	https://github.com/apache/opennlp ; https://opennlp.apache.org/index.html (consultés le 22.12.2021)
OpenTExt EnCase Forensic	Outil	Propriétaire	Guidance Software (OpenText depuis 2017)	1998		Récupération de données; Reconnaissance optique de caractères imprimés; Reconnaissance d'images; Visualisation de chronologie; Garantie de la valeur probante (empreinte digitale numérique); Création d'un rapport synthétique	"OpenText™ EnCase™ Forensic is a court-proven solution for finding, decrypting, collecting and preserving forensic data from a wide variety of devices, while ensuring evidence integrity and seamlessly integrating investigation workflows." (site Opentext)	Krichenbaum 2010, p. 18; Lee 2018; Vinh-Doyle 2017	https://security.opentext.com/encase-forensic (consulté le 21.12.2021)

Nom	Type	Accessibilité	Développeur	Année début	Année fin	Fonctionnalités	Description	Référence(s) bibliographique(s)	Liens
Outil de Constitution et de Traitement Automatisé des Versements Électroniques (OCTAVE)	Outil	Libre	Archives de France			Dédoublonnage; Remaniement de l'arborescence; Identification de format; Renommage de fichiers; Création d'un bordereau d'élimination; Création d'un bordereau de versement; Ajout de métadonnées descriptives intellectuelles; Ajout de métadonnées descriptives techniques; Prévisualisation de fichiers	"Les Archives de France ont travaillé au développement d'un Outil de Constitution et de Traitement Automatisé des Versements Électroniques (OCTAVE). Basé sur le logiciel libre Docuteam Packer, développé en Java, OCTAVE permet à l'archiviste de transformer ses arborescences bureautiques en SIP aux formats SEDA 1 et SEDA 2.1." (site francearchives)	Naud 2019	https://francearchives.fr/fr/article/88482499 ; https://github.com/culturecommunication/octave (consultés le 22.12.2021)
Pattern	Outil	Libre				Analyse des sentiments (NLP)	"Web mining module for Python, with tools for scraping, natural language processing, machine learning, network analysis and visualization." (github Pattern)	Lee 2018	https://github.com/clips/pattern (consulté le 22.12.2021)
Preservica	Outil	Propriétaire	Presservica			Recherche générale; Prévisualisation de fichiers; Conversion de format; Migration de support;	"Preservica's standards-based (OAIS ISO 14721) active preservation software combines all the critical capabilities of successful long-term digital preservation into a single integrated platform. It keeps content safely stored, makes sure it can be found and trusted, provides secure immediate access, and automatically updates files to future-friendly formats." (site Preservica)		https://preservica.com/digital-archive-software-1/active-digital-preservation (consulté le 23.12.2021)
PRONOM Unique Identifiers (PUIDs)	Outil	Libre	UK National Archives	2002		Attribution d'identifiants pérennes	"The PRONOM Unique Identifier (PUID) is an extensible scheme for providing persistent, unique and unambiguous identifiers for records in the PRONOM registry. In the first instance, PUIDs are being assigned to file formats, with over 130 of the most common formats already assigned identifiers, and more being added on a regular basis." (site PRONOM)	Harvey et Thompson 2010	https://www.nationalarchives.gov.uk/aboutapps/pronom/puid.htm (consulté le 25.12.2021)

Nom	Type	Accessibilité	Développeur	Année début	Année fin	Fonctionnalités	Description	Référence(s) bibliographique(s)	Liens
Protégé	Outil	Libre	Stanford Center for Biomedical Informatics Research at the Stanford University School of Medicine	1999		Gestion d'ontologies	"Protégé Desktop suppoProtégé Desktop supports creation and editing of one or more ontologies in a single workspace via a completely customizable user interface. Visualization tools allow for interactive navigation of ontology relationships. Advanced explanation support aids in tracking down inconsistencies. Refactor operations available including ontology merging, moving axioms between ontologies, rename of multiple entities, and more." (site Protégé)	Dobreski et al. 2019	https://protege.stanford.edu/ (consulté le 23.12.2021)
pyLDAvis	Outil	Libre	Ben Mabey			Topic modeling (NLP)	"Python library for interactive topic model visualization." (github pyLDAvis)	Goodman 2019; Hutchinson 2020	https://github.com/bmabey/pyLDAvis (consulté le 22.12.2021)
ReSip	Outil	Libre	Programme interministériel archives numérique (France)	2013		Remaniement de l'arborescence; Ajout de métadonnées descriptives intellectuelles; Ajout de métadonnées techniques; Importation de documents	"L'application ReSIP, basée sur la bibliothèque sedalib, permet de construire et manipuler des structures arborescentes d'archives, d'en éditer les métadonnées, de les importer et exporter sous la forme de SIP, sous la forme de hiérarchie disque ou encore sous forme csv pour les plans de classement." (site vitam)	Naud 2019	http://www.programmevitam.fr/pages/ressources/resip/ (consulté le 22.12.2021)
scikit-learn	Outil	Libre	David Cournapeau	2007		Classification et catégorisation de documents (NLP)	"scikit-learn is a Python module for machine learning built on top of SciPy and is distributed under the 3-Clause BSD license." (github scikit-learn)	Goodman 2019, p. 15	https://github.com/scikit-learn/scikit-learn ; https://scikit-learn.org/stable/ (consultés le 22.12.2021)
ScopeArchiv	Outil	Propriétaire	Scope			Création d'un plan de classement; Ajout de métadonnées descriptives intellectuelles; Ajout de métadonnées descriptives techniques; Contrôle des droits d'accès; Conversion de format; Extraction de	"scopeArchiv™ est une solution informatique d'archivage complète pour les archives privées et publiques. Cette solution standard dotée d'une structure modulaire prend en charge les processus opérationnels types: pré-archivage, transfert et mise à disposition des documents, valorisation et stockage, transmission, recherche et commande de documents archivés sur Internet ou via l'intranet. scopeIngest complète l'offre de scopeArchiv™ pour en faire un système d'informations archivistiques hybride et offre un archivage numérique à long terme conforme aux spécifications de l'OAIS (ISO 14721)." (site Scope)		https://www.scope.ch/fr/apercu-des-produits/scopearchiv/ (consulté le 30.12.2021)

Nom	Type	Accessibilité	Développeur	Année début	Année fin	Fonctionnalités	Description	Référence(s) bibliographique(s)	Liens
						métadonnées; Importation d'un plan de classement préexistant; Importation de documents; Contrôle du vocabulaire; Recherche générale; Recherche géospatiale; Recherche plein-texte; Recherche à facettes; Recherche géospatiale			
Siegfried	Outil	Libre	Richard Lehane	2014		Identification de format	"Siegfried is a signature-based file format identification tool, implementing: the National Archives UK's PRONOM file format signatures; freedesktop.org's MIME-info file format signatures; the Library of Congress's FDD file format signatures (beta); Wikidata (beta)." (github siegfried)	coline1 2019; Hutchinson 2020	https://github.com/richardlehane/siegfried ; https://www.itforarchivists.com/siegfried (consultés le 22.12.2021)
Social Networks and Archival Context (SNAC)	Outil	Libre	University of Virginia Library; National Archives and Records Administration	2010		Contrôle d'autorité; Contrôle du vocabulaire	"SNAC (Social Networks and Archival Context) is a free, online resource that helps users discover biographical and historical information about persons, families, and organizations that created or are documented in historical resources (primary source documents) and their connections to one another. Users can locate archival collections and related resources held at cultural heritage institutions around the world. SNAC is an international cooperative including, but not limited to, archives, libraries, and museums, that is working to build a corpus of reliable descriptions of people, families, and organizations that link to and provide a contextual understanding of historical records." (site SNAC)	Bailey 2013	https://snaccooperative.org/ (consulté le 20.12.2021)
Solr	Outil	Libre	Apache Software Foundation	2006		Reconnaissance optique de caractères imprimés; Recherche plein-texte; Recherche	"Apache Solr is an enterprise search platform written in Java and using Apache Lucene. Major features include full-text search, index replication and sharding, and result faceting and highlighting." (github Solr)	Makhlouf et al. 2020	https://solr.apache.org/ ; https://github.com/apache/solr/ (consultés le 5.1.2022)

Nom	Type	Accessibilité	Développeur	Année début	Année fin	Fonctionnalités	Description	Référence(s) bibliographique(s)	Liens
						géospatiale; Recherche à facettes			
Spark Archives Electronic Edition (SAEE)	Outil	Propriétaire	Spark Archives			Importation de documents; Garantie de la valeur probante (empreinte digitale numérique); Reconnaissance optique de caractères imprimés; Validation de format;Contrôle de signature électronique; Journalisation	"SAEE is a software package, which allows implementation of an Electronic Filing System (EFS) with an evidential value. It integrates functionalities for the capture of natively digital electronic documents, digitised by scanners, Multi-Function Printers (MFP) or smartphones." (site spark archives)		https://www.spark-archives.com/en/SAEE-Spark-Archives-Electronic-Edition (consutlé le 22.12.2021)
StrucTool	Outil	Propriétaire	Archives fédérales suisses	2019		Création d'un plan de classement; Ajout de métadonnées descriptives intellectuelles; Calcul de la valeur archivistique; Création d'un bordereau de versement	"StrucTool est une application permettant d'élaborer, de valider et d'administrer des structures (systèmes de classement et structures libres) ainsi que de créer des bordereaux de versement. [...] L'application StrucTool est utilisée dans le cadre des processus suivants : validation/actualisation (y c. évaluation) de structures (systèmes de classement et structures libres); versement de documents analogiques." (site Archives fédérales suisses, manuel d'utilisation) ; "L'outil StrucTool comporte des fonctionnalités présentes dans différents processus, qui sont décrites dans ce chapitre. Il s'agit des fonctionnalités suivantes : la page d'accueil avec les vues d'ensemble des structures et des versements; les différentes vues; le traitement en masse; les actions sur les éléments (éditer, déplacer, etc.); la recherche; l'importation et l'exportation." (site Archives fédérales suisses, manuel d'utilisation)		https://www.bar.admin.ch/bar/fr/home/gestion-de-l-information/outils-et-instruments/structool--creer-et-gerer-des-structures.html; https://www.bar.admin.ch/dam/bar/fr/dokumente/StrucTool/Benutzerhandbuch.pdf.download.pdf/Benutzerhandbuch.pdf (consultés le 21.12.2021)
Tesseract	Outil	Libre	Google	1985		Reconnaissance optique de caractères imprimés	"Tesseract is an open source text recognition (OCR) Engine, available under the Apache 2.0 license" (site tesseract)		https://github.com/tesseract-ocr/tesseract ; https://tesseract-ocr.github.io/ (consultés le 20.12.2021)

Nom	Type	Accessibilité	Développeur	Année début	Année fin	Fonctionnalités	Description	Référence(s) bibliographique(s)	Liens
Textract	Outil	Libre	Dean Malmgren	2014		Extraction de texte; Extraction de métadonnées	"As undesireable as it might be, more often than not there is extremely useful information embedded in Word documents, PowerPoint presentations, PDFs, etc—so-called “dark data”—that would be valuable for further textual analysis and visualization. While several packages exist for extracting content from each of these formats on their own, this package provides a single interface for extracting content from any type of file, without any irrelevant markup." (site textract)	Hutchinson 2020	https://textract.readthedocs.io/en/stable/index.html (consulté le 20.12.2021)
TextRazor	Outil	Propriétaire	TextRazor Lt	2011		Reconnaissance d'entités nommées (NLP); Utilisation de Linked data; Topic modeling (NLP); Classification et catégorisation de documents (NLP)	"TextRazor uses state-of-the-art Natural Language Processing and Artificial Intelligence techniques to parse, analyze and extract semantic metadata from your content." (site TextRazor)	coline1 2019	https://www.textrazor.com/ (consulté le 20.12.2021)
Tika	Outil	Libre	Apache Software Foundation	2007		Extraction de métadonnées; Identification de langue; Extraction de texte	"The Apache Tika™ toolkit detects and extracts metadata and text from over a thousand different file types (such as PPT, XLS, and PDF). All of these file types can be parsed through a single interface, making Tika useful for search engine indexing, content analysis, translation, and much more." (site tika)	Makhlouf et al. 2020	https://tika.apache.org/ (consulté le 21.12.2021)
Transkribus	Outil	Libre	Read Coop	2013		Reconnaissance optique de caractères manuscrits; Reconnaissance d'images	"Transkribus is a comprehensive platform for the digitisation, AI-powered text recognition, transcription and searching of historical documents – from any place, any time, and in any language." (site Transkribus)	coline1 2019	https://readcoop.eu/transkribus/ ; https://github.com/Transkribus/ (consultés le 20.12.2021)
TreeSize Professional (TSP)	Outil	Propriétaire	JAM Software	1997		Dédoublonnage; Renommage de fichiers; Création d'un rapport synthétique; Analyse de volumétrie	"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. [...] A high degree of	Belovari 2018, p. 61	https://www.jam-software.com/treesize; https://www.jam-software.com/treesize/features.shtml (consultés le 21.12.2021)

Nom	Type	Accessibilité	Développeur	Année début	Année fin	Fonctionnalités	Description	Référence(s) bibliographique(s)	Liens
							automation is enabled by command line parameters and management of scans scheduling directly in a comfortable, graphical user interface." (site Jam Software)		
Vitam (solution logicielle)	Outil	Libre	Programme interministériel archives numérique (France)	2017		Extraction de métadonnées ; Ajout de métadonnées descripives techniques; Création d'un calendrier de conservation; Contrôle de conformité du SIP; Recherche générale; Création de copies d'accès; Création d'une copie de téléchargement; Recherche à facettes; Contrôle des droits d'accès; Gestion d'ontologies; Journalisation; Analyse de volumétrie; Migration de support; Remaniement de l'arborescence; Validation de format; Identification de format; Conversion de format; Garantie de la valeur probante (empreinte digitale numérique); Extraction de métadonnées	"La solution logicielle Vitam permet la prise en charge, la conservation, la pérennisation et la consultation sécurisée de très gros volumes d'archives numériques. Elle assure la gestion complète du cycle de vie des archives et donc la garantie de leur valeur probante. Elle peut être utilisée pour tout type d'archive, y compris pour des documents classifiés de défense." (github vitam)		https://www.programmevitam.fr/pages/logiciel/ ; https://github.com/ProgrammeVitam/vitam (consultés le 23.12.2021)

Nom	Type	Accessibilité	Développeur	Année début	Année fin	Fonctionnalités	Description	Référence(s) bibliographique(s)	Liens
Waikato environment for knowledge analysis (Weka)	Outil	Libre	University of Waikato	1992		Topic modeling (NLP)	"Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization." (site Weka)	Hutchinson 2020	https://waikato.github.io/weka-wiki/ ; https://www.cs.waikato.ac.nz/ml/weka/ (consultés le 22.12.2021)
Web Curator Tool (WCT)	Outil	Libre	British Library; National Library of New Zealand; National Library of the Netherlands (KB-NL)	2006		Moissonnage du web; Ajout de métadonnées descriptives intellectuelles	"The Web Curator Tool is for managing the selective web harvesting process. The tool is designed for use in libraries and other collecting organisations. It supports collection by non-technical users while still allowing complete control of the web harvesting process. The Web Curator Tool was released as open-source software and can be downloaded from GitHub." (site Web Curator Tool)		https://natlib.govt.nz/collections/digital-preservation/ndha-tools-and-resources/web-curator-tool-wct; https://github.com/WebCuratorTool/webcurator-v2-legacy/wiki/Documentation (consulté le 23.12.2021)
WinCatalog 2020	Outil	Propriétaire	OrangeCat Software, LLC	2001		Extraction de métadonnées; Création d'un rapport de récolement; Visualisation de l'arborescence; Création d'arborescence virtuelle; Prévisualisation de fichiers; Dédoublonnage	"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." (site Wincatalog)	Kim 2006, p. 96	https://www.wincatalog.com/; https://www.wincatalog.com/features.html (consultés le 21.12.2021)

Annexe 2 : Tableau synoptique des projets et autres références

Nom	Type	Développeur	Année début	Année fin	Fonction	Description	Référence(s) bibliographique(s)	Liens
AENeas	Projet	Archives de l'État de Neuchâtel (AEN)	2015		Evaluation; Accroissement; Classification; Description; Préservation; Diffusion	Concept d'archivage numérique articulé autour des logiciels constituant la Suite Archi (ArchiClass; ArchiRef; ArchiSelect; ArchiVision; ArchiPeren; ArchiInfo), convrant l'ensemble des fonctions archivistiques. Débuté en 2011, à ce jour seul le premier outil du concept, ArchiClass a été développé.	Makhlouf et. all 2020; Oguey et Schneiter 2018	https://www.ne.ch/autorites/DESC/SCNE/archives-etat/numerique/Pages/70-AENEAS.aspx
Architypes	Projet; Vocabulaire	W3C (Richard Wallis)	2015		Description	"The mission of this group is to discuss and prepare proposal(s) for extending Schema.org schema for the improved representation of digital and physical archives and their contents. The goal being focused upon the creation and future maintenance of an archive.schema.org extension." (site W3, architypes)	Hooland 2018	https://www.w3.org/community/architypes/ (consulté le 23.12.2021)
BagIt	Projet	Library of Congress (LOC)			Préservation; Accroissement	"BagIt, a set of hierarchical file layout conventions for storage and transfer of arbitrary digital content. A "bag" has just enough structure to enclose descriptive metadata "tags" and a file "payload" but does not require knowledge of the payload's internal semantics. This BagIt format is suitable for reliable storage and transfer." (site ietf)		https://datatracker.ietf.org/doc/html/rfc8493 (consulté le 21.12.2021)
Bibframe Lite + archives	Vocabulaire	Zepheira			Description; Diffusion	"Bibframe Archive is a starting point for archival description which builds on the Bibframe Lite vocabulary. It is framework conformant to Bibframe and Describing Archives: A Content Standard (DACS). Where possible, the vocabulary is link-compatible with the US Library of Congress's BIBFRAME vocabulary. Currently, we are analyzing encoding standards including Encoded Archival Description and Encoded Archival Context-Corporate Bodies to expand Bibframe Archive and prototype tools for transforming XML to Linked Data." (site Bibframe)	Matienzo et al. 2017; Dobreski et al. 2019	http://bibfra.me/view/archive/ (consulté le 23.12.2021)

Nom	Type	Développeur	Année début	Année fin	Fonction	Description	Référence(s) bibliographique(s)	Liens
BitCurator Project	Projet	School of Information and Library Science at the University of North Carolina at Chapel Hill (SILS); Maryland Institute for Technology in the Humanities (MITH); BitCurator Consortium	2011	2014	Evaluation; Classification; Description;	"The BitCurator NLP project is developing software for collecting institutions to extract, analyze, and produce reports on features of interest in text extracted from born-digital materials contained in collections. We are using open source natural language processing libraries to identify items likely to be relevant to preservation, information organization, and access activities. These may include entities (e.g. persons, places, and organizations), potential relationships among entities (e.g. those entities that appear together within documents or set of documents), and topic models to provide insight into how concepts are naturally clustered within the documents. We are developing software that will allow users to create customized reports from text discovered in disk images, providing both command-line executables and a public Python API to extend the capabilities of external tools." (site BitCurator)	Hutchinson 2020; Lee 2018; Shein 2014	https://bitcurator.net/bitcurator/ (consulté le 23.12.2021)
Born Digital Collections : An Inter-Institutional Model for Stewardship (AIMS)	Projet	University of Virginia Libraries; Stanford University Libraries Academic Resources; University of Hull Library : Yale University Library; Andrew W. Mellon Foundation (support)	2009	2011	Accroissement; Classement; Description; Diffusion	"Into this climate, the AIMS partners proposed an inter-institutional framework for stewarding born-digital content. The AIMS partners realized that they could not solve all problems associated with born-digital materials but decided to focus their attention on professional practice defined by archival principles and by the current state of collections at the partner institutions." (AIMS Final text)	Bailey 2013; Shein 2014	https://dcs.library.virginia.edu/files/2013/02/AIMS_final_text.pdf (consulté le 23.12.2021)
CIDOC CRM	Modèle	ICOM/CIDOC Standard Group	2006		Description; Diffusion	Modèle conceptuel (ISO 21127:2006). "The CIDOC CRM represents an 'ontology' for cultural heritage information i.e. it describes in a formal language the explicit and implicit concepts and relations relevant to the documentation of cultural heritage. The primary role of the CIDOC CRM is to serve as a basis for mediation of cultural heritage information and thereby provide the semantic 'glue' needed to transform today's disparate, localised information sources into a coherent and valuable global resource." (site CIDOC CRM)	Gracy 2015, p. 256	http://www.cidoc-crm.org/ (consulté le 19.12.2021)

Nom	Type	Développeur	Année début	Année fin	Fonction	Description	Référence(s) bibliographique(s)	Liens
Civil War Data 150	Projet	The Archives of Michigan; The Internet Archives; Freebase; Digital Scholarship Lab (University of Richmond); Historypin	2011	2013	Description; Diffusion	"CWD150 is exploring the use of Linked Open Data within libraries, archives and museums, and extending the usability and availability of structured data." ConflictHistory.com utilizes the Freebase API and Google Maps Flash to present a dynamic view of the history of war. Live Tweeting the Civil War + 150." (site Civil War Data 150) "The Civil War Data 150 Project was started as a proof of concept to explore the potential of Linked Open Data in Libraries, Archives, and Museums, which has become known generally as #LODLAM. An unsuccessful grant application to the NEH Office of Digital Humanities lead to the first International LODLAM Summit in June of 2011. Right after that I joined the Historypin team full time, and this project has not gotten much love since then." (site Civil War Data 150)	Gracy 2015, p. 251	https://www.civilwardata150.net/ (consulté le 19.12.2021)
Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval (CASPAR)	Projet	The Science and Technology Facilities Council	2006	2009	Préservation	"CASPAR will research, implement, and disseminate innovative solutions for digital preservation based on the OAIS reference model (ISO:14721:2003). (site CASPAR)	Harvey et Thompson 2010; Ross 2012, p. 59	https://www.casparpreserves.eu/ (consulté le 21.12.2021)
Digital Forensics and Born-Digital Content in Cultural Heritage Collections	Projet	Council on Library and Information Resource (CLIR)	2010		Évaluation; Description	"This report introduces the field of digital forensics in the cultural heritage sector and explores some points of convergence between the interests of those charged with collecting and maintaining born-digital cultural heritage materials and those charged with collecting and maintaining legal evidence." (site clir)	Lee 2018	https://www.clir.org/pubs/reports/pub149/
Digital Library Reference Model	Modèle	DELOS NoE (2000 - 2003); DL Org Project (2008 - 2011)	2000	2011	Description; Diffusion	"DL.org Booklets based on core set of outputs: Digital Library Manifesto, Digital Library Checklist, Digital Library Cookbook, Digital Library Reference Model - In a Nutshell. The outputs stem from synergies with international experts and engagement with stakeholder community, professionals, educationalists and students at various stages in their academic careers mainly from the Library and Information space but also from Computer Science. Many of our stakeholders have shared priorities and views through position statements." (site DL Org)	Harvey et Thompson 2010; Ross 2012, p. 50	http://www.dlorg.eu/ (consulté le 22.12.2021)

Nom	Type	Développeur	Année début	Année fin	Fonction	Description	Référence(s) bibliographique(s)	Liens
Digital Preservation Handbook	Guide	Digital Preservation Coalition	2001		Préservation	"We enable our members to deliver resilient long-term access to digital content and services, helping them to derive enduring value from digital assets and raising awareness of the strategic, cultural and technological challenges they face. We achieve our aims through advocacy, community engagement, workforce development, capacity-building, good practice and good governance." (site DPC)	Ross 2012	https://www.dpconline.org/ (consulté le 19.12.2021)
Digital Preservation Policy	Projet	National Archives of Australia (NAA)			Préservation	"Digital preservation aims to address the following risks: The content of digital records becomes inaccessible due to future software obsolescence, data loss due to the obsolescence or failure of the hardware or media used to store digital records, data loss due to inadvertent or malicious alteration of content, and inauthentic or unreliable data due to incomplete or inadequate capture of digital records and metadata at the time of transfer. "This Policy describes the digital archiving principles and approaches adopted by the National Archives of Australia (Archives) to ensure these risks are mitigated as much as possible. Further policy documents, procedures, standards, and guidance will be developed in future to address specific aspects of the Policy. This Policy addresses the following target groups: Archives' staff, Commonwealth Government agencies, Commonwealth Government agencies, Expert groups in the digital archiving community, Public clients." (site NAA)	Makhlouf Shabou 2015, p. 202; Rolan et al. 2019, p. 194	https://www.naa.gov.au/about-us/our-organisation/accountability-and-reporting/archival-policy-and-planning/digital-preservation-policy#standards (consulté le 21.12.2021)
EDINA	Projet	University of Edinburgh			Description	"EDINA are specialists in developing and delivering digital products including large-scale online services, mobile apps and digital tools for education. Our expertise includes geospatial and satellite data, computational learning, text and data mining, e-preservation and machine learning." (site EDINA) "Edina Unlock Places (API to allow the user to match a placename string against one or more gazetteers. The geoparser uses this to find candidate locations for placename strings extracted by the geotagging step. The API provides a simple and flexible way to search different gazetteers, with results available in a range of formats.), Edina Unlock Texxt (It allows you to submit complete texts for geoparsing, either individually or in bundles. It is, in effect, an online version of the geoparser pipeline and is probably the simplest way of using the geoparser, if no customisation is required. There are subtle and	Clough 2010, p. 139	https://edina.ac.uk/ (consulté le 19.12.2021)

Nom	Type	Développeur	Année début	Année fin	Fonction	Description	Référence(s) bibliographique(s)	Liens
						unavoidable differences between the Unlock version and the downloadable geoparser package, meaning that results will not necessarily be identical. If your needs are complex it may be better to install a local copy of the geoparser.)" (site EDINA)Digimap is an online map and data delivery service, available by subscription to UK Higher and Further Education establishments. Operated by EDINA at the University of Edinburgh, Digimap offers a number of data collections, including Ordnance Survey, historical, geological, LiDAR and marine maps and spatial data." (site EDINA) "Digimap offers access to a range of datasets for the purposes of education and research. Create or interrogate a map online by selecting an appropriate base map, adding annotations and customising the content, use measurement and query tools to learn more about any study area. Download the raw spatial data in a wide range of formats for use in local GIS, CAD or image processing software." (site EDINA)		
Electronic Resource Preservation and Access Network (ERPANET)	Projet	The Humanities Advanced Technology and Information Institute (HATII) (University of Glasgow); Nationaal Archief van Nederland; Institute for Archival and Library Science; Istituto di Studi per la Tutela dei Beni Archivistici e Librari (Università degli Studi di Urbino Carlo Bo); Archives fédérales suisses (AFS)	2001	2005	Préservation	"ERPANET will bring together memory organisations (museums, libraries and archives), ICT and software industry, research institutions, government organisations (including local ones), entertainment and creative industries, and commercial sectors (including for example pharmaceuticals, petro-chemical, and financial). The dominant feature of ERPANET will be the provision of a virtual clearinghouse and knowledge-base on state-of-the-art developments in digital preservation and the transfer of that expertise among individuals and institutions." (site ERPANET)	Ross 2006	https://www.erpanet.org/index.php (consulté le 19.12.2021)
Encoded Archival Context - Corporate bodies, Persons, and Families (EAC CPF)	Vocabulaire	Instituto per i beni artistici culturali et naturali delle Regione Emilia-Romagna			Description	"Ontology that describes all the elements and the attributes of the XML schema and refers to tag library and xml diagram for technical specification. This beta version of the ontology is now available as rdf file and as a graph principally to support the archivist's work during the description of individual or organization responsible for the creation, accumulation, or assembly the described materials. This first result was added at the W3C document on vocabulary and dataset used in library and archives of the Library Linked Data working group." (site EAC CPF)	Gracy 2015, p. 254	http://archivi.ibc.regione.emilia-romagna.it/ontology/reference_document/referencedocument.html

Nom	Type	Développeur	Année début	Année fin	Fonction	Description	Référence(s) bibliographique(s)	Liens
Europeana Data Model	Modèle	Europeana (UE)	2005		Description; Diffusion	"Our Mission. Europeana empowers the cultural heritage sector in its digital transformation. We develop expertise, tools and policies to embrace digital change and encourage partnerships that foster innovation. We make it easier for people to use cultural heritage for education, research, creation and recreation. Our work contributes to an open, knowledgeable and creative society. Our vision. Europeana imagines a cultural heritage sector powered by digital and a Europe powered by culture, giving it a resilient, growing economy, increased employment, improved well-being and a sense of European identity. Who we are. Europeana is made possible by the collaboration of three interlinked expert organisations who share the vision of a cultural heritage sector transformed by digital, and a Europe transformed by culture. Together, we call this ecosystem the Europeana Initiative." (site Europeana)	Gracy 2015, p. 251; Matienzo et al. 2017, p. 112	https://pro.europeana.eu/ (consulté le 19.12.2021)
Exploring Collaborations to Harness Objects with a Digital Environment for Preservation (ECHO DEPository Project)	Projet	University of Illinois at Urbana-Champaign Library, Graduate School of Library and Information Science	2004	2010	Préservation; Diffusion	"The ECHO DEPository project pulled together several streams of activities aimed at helping to answer the question of how digital resources will be identified, archived, and preserved for the future. ECHO DEPository explored ways for libraries and repositories to share and preserve digital information in a variety of formats, including Web-based government publications, historical documents and photos, sound and video recordings, websites and other digital resources. Partners collaborated to produce tools, practices, evaluations and research that will help in selecting and preserving electronic resources in a variety of digital repositories." (site digitalpreservation)	Harvey et Thompson 2010, p. 318	https://www.digitalpreservation.gov/partners/echodep.html (consulté le 22.12.2021)
Forgotten History	Projet	Arton Foundation	2016		Description; Diffusion	"The www.forgottenheritage.eu database was created in 2018 as part of the "Forgotten Heritage: European Avant-garde Art Online" project, an initiative by the Arton Foundation (www.fundacjaarton.pl) within the Creative Europe programme. The project was realized in collaboration with the Luca School of Arts (https://www.luca-arts.be/nl), the Office for Photography (croatian-photography.com) and the KUMU Art Museum (https://kumu.ekm.ee/en/), with the goal of digitizing and making available online works by avant-garde artists from Poland, Croatia, Belgium and Estonia, with a focus on the 1960s and 1970s. Within the initiative, the partners conducted research in numerous archives, both private and institutional, then studied and arranged		https://www.forgottenheritage.eu/ (consulté le 22.12.2021)

Nom	Type	Développeur	Année début	Année fin	Fonction	Description	Référence(s) bibliographique(s)	Liens
						them, and made them available in a database designed and created by the Plural collective." (site Forgotten Heritage)		
Future Proof	Projet	New South Wales (NSW) State Archives and Records	2007	2018	Classement; Préservation	"State Archives and Records Authority NSW seeks to ensure that the people and Government in NSW have ready access to records which illuminate history, enrich the life of the community and support good and accountable government. Digital records of government are particularly vulnerable to degradation, alteration and loss through time." (site Future Proof)	Rolan et al. 2019, p. 190	https://futureproof.records.nsw.gov.au/ (consulté le 20.12.2021)
futureArch project	Projet	Bodleian Library; The Andrew W. Mellon Foundation (support)	2008	2012	Préservation; Diffusion	"In a nutshell, the project's objective is to provide the Bodleian with infrastructure, policy, processes and skills to curate hybrid archives. [...] To that end, futureArch is integrating curation of born-digital archives into the everyday work of the Library's Western Manuscripts department and developing Bodleian Electronic Archives & Manuscripts' (BEAM) as an infrastructure for curating and disseminating the born-digital elements of hybrid archives while maintaining links with the traditional elements of collections to maintain context" (site digital humanities, futurearch-project)	AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship 2012, p. IV	https://digital.humanities.ox.ac.uk/project/futurearch-project; https://futurearchives.blogspot.com/ (consultés le 22.12.2021)
Genève Archivage à Long Terme d'Archives Électroniques (Gal@tae)	Projet	Archives d'État de Genève	2011	2013	Versement; Préservation	"L'objectif de ce projet était de mettre en place tous les éléments permettant un archivage électronique pérenne des documents, de la création dans un système d'information métier de paquets de données électroniques à archiver au dépôt sur la plate-forme de pérennisation des Archives fédérales suisses, en passant par les étapes de transfert et de contrôle de qualité aux Archives d'Etat. Ce projet pilote a permis de développer les compétences en matière d'archivage électronique pour les Archives d'Etat et les informaticiens, de mettre en place les procédures et processus permettant d'assurer cet archivage pérenne et de tester la solution technique" (Anouk Dunant Gonzenbach, 2013)	Makhlouf 2015, pp. 201-202	https://ge.ch/archives/actualites/archives-detat-conservent-desormais-donnees-nees-numeriques; http://www.ressi.ch/num14/article_93 (consultés le 22.12.2021)

Nom	Type	Développeur	Année début	Année fin	Fonction	Description	Référence(s) bibliographique(s)	Liens
GeoNames	Vocabulaire	Marc Wick; Christophe Boutreux			Description; Diffusion	"The GeoNames Ontology makes it possible to add geospatial semantic information to the Word Wide Web. All over 11 million geonames toponyms now have a unique URL with a corresponding RDF web service. Other services describe the relation between toponyms." (site GeoNames)	Gracy 2015, p. 269	http://www.geonames.org/ (consulté le 19.12.2021)
Georeferencer	Projet	British Library	2011		Description; Diffusion	"Georeferencing involves assigning points on a map image to corresponding geographical coordinates. It links the map to its spatial location on the ground using universal geographic standards (latitude / longitude)." (site British Library)		https://www.bl.uk/projects/georeferencer (consulté le 19.12.2021)
Guidelines for the Preservation of Digital Heritage	Modèle	UNESCO	2003		Préservation	"These Guidelines form a small part of a far-seeing campaign by UNESCO to improve access to digital heritage for all the world's peoples, and to ensure that the means of preserving their digital heritage are in the hands of every community." (site UNESCO)	Ross 2012, p. 45	http://www.unesco.org/new/en/communication-and-information/resources/publications-and-communication-materials/publications/full-list/guidelines-for-the-preservation-of-digital-heritage/ (consulté le 21.12.2021)
InterPARES	Projet	School of Library, Archival and Information Studies (University of British Columbia)	1999		Préservation	"The International Research on Permanent Authentic Records in Electronic Systems (InterPARES) aims at developing the knowledge essential to the long-term preservation of authentic records created and/or maintained in digital form and providing the basis for standards, policies, strategies and plans of action capable of ensuring the longevity of such material and the ability of its users to trust its authenticity. The findings and products of the first three phases of the project can be found on this website." (site InterPARES)	Harvey et Thompson 2010	http://www.interpares.org/ (consulté le 21.12.2021)
IRI Guidelines	Modèle	Australian Government Department of Finance (The Australian Government Linked Data Working Group)	2012		Description; Diffusion	"As Linked Data technologies advance and become commonplace, it will be necessary for Government to become responsive to the demands of its citizens, as well as its own entities, with regards to Linked Data's use. Developing government arrangements and establishing technical mechanisms for Linked Data implementations will ensure Australian individuals, businesses and organisations can benefit from the opportunities these technologies offer. The Group is a community of Commonwealth Government Linked Data experts and champions with invited, non-voting, participation of individuals, corporations and other entities. In addition to drafting policy and technical guidance on the implementation of Linked Data for the Australian	Rolan et al. 2019, p. 194	https://www.linked.data.gov.au/ (consulté le 22.12.2021)

Nom	Type	Développeur	Année début	Année fin	Fonction	Description	Référence(s) bibliographique(s)	Liens
						Government, members of the group also supply some core technical Linked Data services." (site Australian Government Linked Data Working Group)		
Library Linked Data Incubator Group (LLD)	Projet	W3C		2011	Description	"The mission of the Library Linked Data incubator group is to help increase global interoperability of library data on the Web, by bringing together people involved in Semantic Web activities—focusing on Linked Data—in the library community and beyond, building on existing initiatives, and identifying collaboration tracks for the future. The group has explored how existing building blocks of librarianship, such as metadata models, metadata schemas, standards and protocols for building interoperability and library systems and networked environments, encourage libraries to bring their content, and generally re-orient their approaches to data interoperability towards the Web, also reaching to other communities. It has also envisioned these communities as a potential major provider of authoritative datasets (persons, topics...) for the Linked Data Web. As these evolutions raise a need for a shared standardization effort within the library community around (Semantic) Web standards, the group has sought to refine the knowledge of this need, express requirements for standards and guidelines, and propose a way forward for the library community to contribute to further Web standardization actions." (site LLD)	Gracy 2015, p. 257	https://www.w3.org/2005/Incubator/lld/ (consulté le 21.12.2021)

Nom	Type	Développeur	Année début	Année fin	Fonction	Description	Référence(s) bibliographique(s)	Liens
Library of Congress Name Autority File (LCNAF)	Vocabulaire	Library of Congress (LOC)	2011		Description; Diffusion	The Library of Congress Name Authority File (NAF) file provides authoritative data for names of persons, organizations, events, places, and titles. Its purpose is the identification of these entities and, through the use of such controlled vocabulary, to provide uniform access to bibliographic resources. Names descriptions also provide access to a controlled form of name through references from unused forms, e.g. a search under: Snodgrass, Quintus Curtius, 1835-1910 will lead users to the authoritative name for Mark Twain, which is, "Twain, Mark, 1835-1910." Names may also be used as subjects in bibliographic descriptions, so they may be combined with controlled values from subject heading schemes, such as LCSH. Library of Congress Names includes over 8 million descriptions created over many decades and according to different cataloging policies. LC Names is officially called the NACO Authority File and is a cooperative effort in which participants follow a common set of standards and guidelines. (site web)	Gracy 2015. p. 256, 270	https://id.loc.gov/authorities/names.html (consulté le 21.12.2021)
libratom	Vocabulaire	State Archives of North Carolina; the School of Information and Library Science at UNC Chapel Hill.	2019		Topic modeling (NLP); Création d'un rapport synthétique; Reconnaissance d'entités nommées (NLP)	"[libratom] focuses on the development of a toolset to scan PST and MBOX email sources, produce reports describing content and metadata, and apply NLP to extract and categorize entities discovered in message content. These features are exported in a clearly documented SQLite database schema to support data analytics and machine learning tasks." (site RATOM)	Higgs 2019; Hutchinson 2020	https://github.com/libratom/libratom ; https://ratom.web.unc.edu/tools-and-code/ (consultés le 20.12.2021)
Linked Jazz	Projet	LOD (Semantic Lab at Pratt Institute School of Library Information Science)	2011		Description; Diffusion	"To create LOD, we developed a suite of tools: a transcript analyzer, a name mapping and curator tool, and a crowdsourcing tool. These tools operate together to find names mentioned during the interview in order to assign a positive identification to each, disambiguating names using online resources like DBpedia and VIAF. The transcript analyzer also recognizes the question and answer structure of the oral history. As people are mentioned by an interviewee, simple RDF triples between interviewee and persons mentioned are created in the form of knowsOf. These triples can then be mapped to the correlating block." (site Linked Jazz)	Gracy 2015, p. 251	https://linkedjazz.org/ (consulté le 19.12.2021)

Nom	Type	Développeur	Année début	Année fin	Fonction	Description	Référence(s) bibliographique(s)	Liens
Linked Open Capac and Archives Hub (LOCAH)	Projet	Jisc; United Kingdom Office for Library and Information Networking (UKOLN); University of Manchester	2011	2012	Description; Diffusion	"The Archives Hub is a free online service giving access to descriptions of archives held in UK repositories (such as universities, company archives and local history centres). It does not hold any archive material itself but provides a means to cross-search archival descriptions from different institutions.It also provides descriptions of online resources, often including digital content, and holds information on individual repositories." (site LOCAH)	Gracy 2015, p. 251	https://archiveshub.jisc.ac.uk/ (consulté le 20.12.2021)
Linking Lives	Projet	Archives Hub			Description	"Linking Lives explored ways to present Linked Data. We aimed to show that archives can benefit from being presented as a part of the diverse data sources on the Web to create full biographical pictures, enabling researchers to make connections between people and events. Linking Lives built upon the Locah project. Locah was a JISC-funded project to expose the Archives Hub descriptions as Linked Data." (site Linking Lives)	Matienzo et al. 2017, p. 112	http://linkinglives.archiveshub.ac.uk/ (consulté le 20.12.2021)
Linking Open Description of Events (LODE)	Vocabulaire	Ryan Shaw (School of Information and Library Science at the University of North Carolina at Chapel Hill); Lynda Hardman (Centrum Wiskunde & Informatica (CWI)l; Raphael Troncy (EURECOM Graduate School & Research Center)	2009		Description; Diffusion	"Ontology for publishing descriptions of historical events as Linked Data, and for mapping between other event-related vocabularies and ontologies." (site LODE)	Gracy 2015, p. 257	https://linkedevents.org/ontology/ (consulté le 19.12.2021)
LOD-LAM	Projet		2011		Diffusion	"We look into the history of LAMs; how they started and evolved over the years. We also focus on their increasing collaboration to offer integrated functions. At LOD-LAM, we've invested in various resources to help us offer authoritative information on LAMs. Whether general information on LAMs, or something specific on a particular library, archive or museum, we've it covered." (site LOD-LAM)	Dobreski et al. 2019, p. 2	http://lod-lam.net/ (consulté le 19.12.2021)

Nom	Type	Développeur	Année début	Année fin	Fonction	Description	Référence(s) bibliographique(s)	Liens
Lucene	Vocabulaire	Apache Software Foundation	2000		Description; Diffusion	"Apache Lucene is a high-performance, full-featured text search engine library written in Java." (github Lucene)	Makhlouf et al. 2020	https://github.com/apache/lucene ; https://lucene.apache.org/ (consultés le 21.12.2021)
Martha Ballard's Diary	Projet	Cameron Blevins (University of Colorado Denver)	2009		Description; Diffusion	"In A Midwife's Tale, Laurel Ulrich describes the challenge of analyzing Martha Ballard's exhaustive diary, which records daily entries over the course of 27 years: "The problem is not that the diary is trivial but that it introduces more stories than can be easily recovered and absorbed." (25) This fundamental challenge is the one I've tried to tackle by analyzing Ballard's diary using text mining. There are advantages and disadvantages to such an approach – computers are very good at counting the instances of the word "God," for instance, but less effective at recognizing that "the Author of all my Mercies" should be counted as well. The question remains, how does a reader (computer or human) recognize and conceptualize the recurrent themes that run through nearly 10,000 entries. One answer lies in topic modeling, a method of computational linguistics that attempts to find words that frequently appear together within a text and then group them into clusters. I was introduced to topic modeling through a separate collaborative project that I've been working on under the direction of Matthew Jockers (who also recently topic-modeled posts from Day in the Life of Digital Humanities 2010)." (site Martha Ballard's Diary)		http://historying.org/martha-ballards-diary/ (consulté le 20.12.2021)
Old Bailey Online	Projet	Digital Humanity Institute (University of Sheffield)			Description; Diffusion	"The Old Bailey Proceedings Online makes available a fully searchable, digitised collection of all surviving editions of the Old Bailey Proceedings from 1674 to 1913, and of the Ordinary of Newgate's Accounts between 1676 and 1772. It allows access to over 197,000 trials and biographical details of approximately 2,500 men and women executed at Tyburn, free of charge for non-commercial use." (site Old Bailey Online) "In addition to the text, accessible through both keyword and structured searching, this website provides digital images of all 190,000 original pages of the Proceedings, 4,000 pages of Ordinary's Accounts, advice on methods of searching this resource, information on the historical and legal background to the Old Bailey court and its Proceedings, and descriptions of published and manuscript materials relating to the trials covered.	Bailey 2013	https://www.oldbaileyonline.org/static/Project.jsp (consulté le 20.12.2021

Nom	Type	Développeur	Année début	Année fin	Fonction	Description	Référence(s) bibliographique(s)	Liens
						Contemporary maps, and images have also been provided." (site Old Bailey Online)		
OpenAIRE	Guide	Digital Curation Centre (DCC)	2004		Préservation	"The Digital Curation Centre (DCC) was launched in March 2004 with support from Jisc and the Engineering and Physical Sciences Research Council (EPSRC) to help solve digital curation and longer-term preservation challenges that could not be tackled effectively by any single institution or discipline. [...] We have a focus on enabling research data management and increasingly around supporting the production and use of data that are findable, accessible, interoperable and reusable (FAIR)." (site OpenAIRE)	Ross 2012	https://www.dcc.ac.uk/about (consulté le 20.12.2021)
PERSIST	Projet	UNESCO	2013		Préservation	"In close cooperation with other stakeholders, PERSIST project has contributed to the sustainability of the information society through long-term preservation and access to information.The objective has been to secure effective mechanisms for governance and access to knowledge and information over long time. To accomplish this, PERSIST has worked on cooperation and dialogue between governments, social organizations, and the IT-industry, and has promoted practical solutions for sustainable digital preservation." (site PERSIST)		https://unescopersist.org/publications/ (consulté le 20.12.2021)
PLANET	Projet	Open Preservation Foundation	2010		Préservation	"The OPF was established in 2010 to sustain the outcomes of an EU-funded research and development project called Planets. One of the first organisations of its kind, the Foundation's goal was to build on the project's collaborative, practical approach to advancing digital preservation." (site PLANET) Produits : JHOVE (open source, extensible software framework for identification, validation, and characterisation); veraPDF (PDF/A validator,	Harvey et Thompson 2010; Ross 2012	https://openpreservation.org/ (consulté 20.12.2021)

Nom	Type	Développeur	Année début	Année fin	Fonction	Description	Référence(s) bibliographique(s)	Liens
						covering all parts of the PDF/A standards); jypylizer (validator and feature extractor for JP2 images); fido (identify the file formats of digital objects)		
PREMIS	Vocabulaire	Library of Congress (LOC)	2005		Description; Diffusion	The PREMIS Data Dictionary for Preservation Metadata is the international standard for metadata to support the preservation of digital objects and ensure their long-term usability. Developed by an international team of experts, PREMIS is implemented in digital preservation projects around the world, and support for PREMIS is incorporated into a number of commercial and open-source digital preservation tools and systems. The PREMIS Editorial Committee coordinates revisions and implementation of the standard, which consists of the Data Dictionary, an XML schema, and supporting documentation. (site premis)		http://www.loc.gov/standards/premis/ ; https://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf (consultés le 22.12.2021)
Presidential Electronic Records PilotSystem (PERPOS)	Projet	Georgia Tech Institute; NARA	2007		Accession, arrangement, preservation, review, description, and creation of finding aids are supported.	"The Presidential Electronic Records PilOt System (PERPOS) is a research prototype for investigating advanced tec hnologies supporting archival decisions in processing Presidential e-records." (Carter et al. 2007)	Lee 2018; Hutchinson 2020; Underwood and Laib 2007)	
PRONOM	Vocabulaire	UK National Archives			Description; Diffusion	"PRONOM is an on-line information system about data file formats and their supporting software products. Originally developed to support the accession and long-term preservation of electronic records held by the National Archives, PRONOM is now being made available as a resource for anyone requiring access to this type of information." (site PRONOM)	Harvey et Thompson 2010	https://www.nationalarchives.gov.uk/PRONOM/ (consulté le 20.12.2021)

Nom	Type	Développeur	Année début	Année fin	Fonction	Description	Référence(s) bibliographique(s)	Liens
Records in Contexts	Projet	Conseil international des Archives (ICA)	2016		Description	"Depuis 2012, l'EGAD, formé de membres venant de quinze pays, développe, en se fondant sur les principes archivistiques, une nouvelle norme pour la description des documents. Pendant ce travail, l'EGAD a pris en compte certaines critiques des pratiques actuelles, les modèles conceptuels nationaux éprouvés ou émergents, les modèles des communautés professionnelles alliées et les possibilités offertes par les technologies de la communication récentes ou émergentes. L'objectif de RiC est à la fois de réconcilier et d'intégrer les quatre normes existantes (la Norme générale et internationale de description archivistique ISAD(G), la Norme internationale sur les notices d'autorité utilisées pour les archives relatives aux collectivités, aux personnes ou aux familles ISAAR(CPF), la Norme internationale pour la description des fonctions ISDF et la Norme internationale pour la description des institutions de conservation des archives ISDIAH) et de s'appuyer sur elles pour aller plus loin." (site Records in Contexts)	Popovici 2019; Hooland 2018	https://www.ica.org/fr/records-in-contexts-modele-conceptuel (consulté le 20.12.2021)
Records In the Cloud (RiC)	Projet	University of British Columbia (UBC) School of Library, Archival and Information Studies, the Faculty of Law, and the Sauder School of Business; the University of Washington School of Information; the University of North Carolina at Chapel Hill School of Information and Library Science; the Mid-Sweden University Department of Information Technology and Media; the University of Applied Sciences of Western Switzerland School of Business Administration; the Cloud Security Alliance, supported by a Social Sciences and Humanities	2012	[2016]	Préservation	"Objectives : to identify and examine in depth the management, operational, legal, and technical issues surrounding the storage and management of records in the cloud; to determine what policies and procedures a provider should have in place for fully implementing the records/archives management regime of the organization outsourcing the records, for responding promptly to its needs, and for detecting, identifying, analyzing and responding to incidents; and to develop guidelines to assist organizations in assessing the risks and benefits of outsourcing records/archives storage and processing to a cloud provider, for writing contractual agreements, certifications and attestations, and for the integration of outsourcing with the organization's records management and information governance programs." (site Records in the Cloud)	Makhlouf 2015, p. 206	http://www.recordsinthecloud.org/ (consulté le 22.12.2021)

Nom	Type	Développeur	Année début	Année fin	Fonction	Description	Référence(s) bibliographique(s)	Liens
		Research Council of Canada (SSHRC) Insight Grant.						
Review, Appraisal, and Triage of Mail (RATOM) Project	Projet	State Archives of North Carolina; the School of Information and Library Science at UNC Chapel Hill.	2019		Evaluation	"The Review, Appraisal, and Triage of Mail (RATOM) project is developing software to assist archives and other collecting organizations with email analysis, selection, and appraisal tasks. The project extends the email processing capabilities currently present in the TOMES software and BitCurator environment, developing additional modules for these tools along with select standalone software to support more advanced workflows." (site RATOM)	Higgs 2019; Hutchinson 2020	https://ratom.web.unc.edu/
Schema Architypes Community Group	Projet	W3C	2015		Description	The mission of this group is to discuss and prepare proposal(s) for extending Schema.org schema for the improved representation of digital and physical archives and their contents. The goal being focused upon the creation and future maintenance of an archive.schema.org extension. (site Architypes)	Matienzo et al. 2017	https://www.w3.org/community/architypes/

Nom	Type	Développeur	Année début	Année fin	Fonction	Description	Référence(s) bibliographique(s)	Liens
spaCy	Vocabulaire	Matt Honnibal	2015		Description; Diffusion	"spaCy is a library for advanced Natural Language Processing in Python and Cython. It's built on the very latest research, and was designed from day one to be used in real products. spaCy comes with pretrained pipelines and currently supports tokenization and training for 60+ languages. It features state-of-the-art speed and neural network models for tagging, parsing, named entity recognition, text classification and more, multi-task learning with pretrained transformers like BERT, as well as a production-ready training system and easy model packaging, deployment and workflow management. spaCy is commercial open-source software, released under the MIT license." (github spaCy)	Hutchinson 2020; Lee 2018	https://spacy.io/ ; https://github.com/explosion/spaCy (consultés le 20.12.2021)
The Personal Archives Accessible in DIGital Media (PARADIGM)	Projet	Bodleian Libraries (University of Oxford); John Rylands University Library (University of Manchester)	2005	2007	Versement; Préservation	"The Personal Archives Accessible in Digital Media (paradigm) project saw the major research libraries of the Universities of Oxford and Manchester come together to explore the issues involved in preserving digital private papers through gaining practical experience in accessioning and ingesting digital private papers into digital repositories, and processing these in line with archival and digital preservation requirements. The project outcomes included: A template for ensuring long-term access for institutional holdings of digital personal papers Best-practice guidelines in the form of a workbook on issues relating to the archiving of personal papers in digital form, made available in sections as they are completed Strengthened local institutional capacity for digital preservation Developed templates for institutional policies for collection development, retention, and disposal Practical test of digital repository softwares DSpace and Fedora and related tools Research resources for 20th century political history in the form of new archival collections." (site Paradigm)	Bailey 2013	http://www.paradigm.ac.uk/; https://ora.ox.ac.uk/objects/uuid:116a4658-deff-4b06-81c5-c9c2071bc6d0 (consultés le 22.12.2021)
Traces through time	Projet	UK National Archives	2014		Description; Diffusion	"A project by The National Archives which combines historical data sets and the latest technology to help researchers find linked records across our collections. Starting with service records from the First World War, the project has so far identified and published over half a million links. This	Bell & Ranade 2015	https://media.nationalarchives.gov.uk/index.php/traces-time-new-tool-finding-linked-records-across-collections/ (consulté le 21.12.2021)

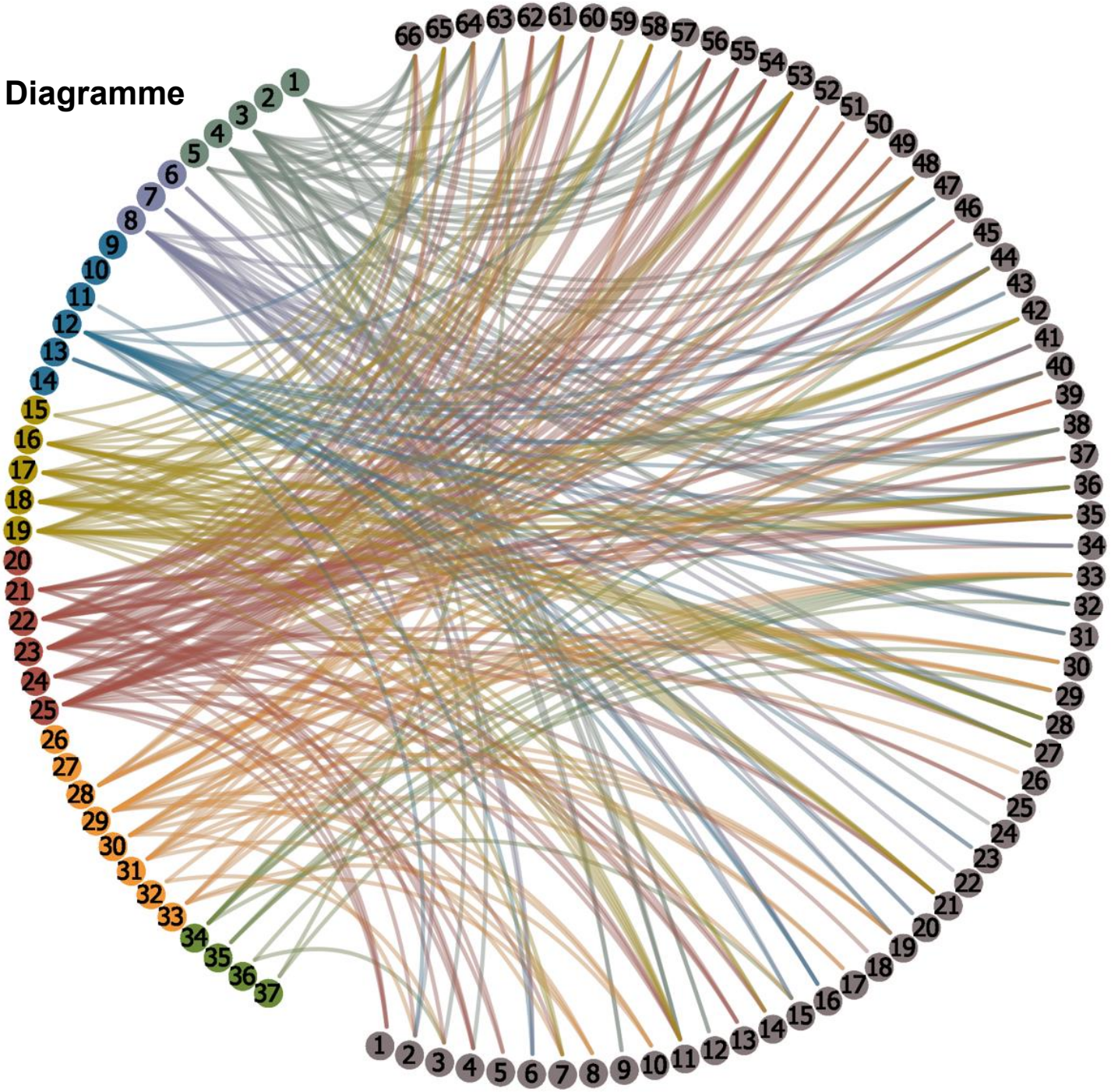
Nom	Type	Développeur	Année début	Année fin	Fonction	Description	Référence(s) bibliographique(s)	Liens
						work enables new insights from archival records and allows people's stories to emerge from the data." (site Traces through time)		
Transforming Online Mail with Embedded Semantics (TOMES)	Projet	State Archives of North Carolina; Utah Division of Archives and Records Service; Kansas Historical Society	2015	2018	Acquisition; Diffusion; Préservation	"The Transforming Online Mail with Embedded Semantics (TOMES) project, generously funded by the National Historical Publications and Records Commission seeks to identify email accounts of public officials with enduring value in order to capture, preserve and provide access to important government records." (site Tomes)	Higgs 2019; Hutchinson 2020	https://www.ncdcr.gov/resources/records-management/tomes (consulté le 22.12.2021)
UCIspace @ the Libraries	Projet	Online Archives of California (University of California (UCI) Libraries)			Description; Diffusion	"UCIspace @ the Libraries is one of a suite of digital scholarship services offered by the the UCI Libraries. To learn more about our other digital scholarship services, please see our website, or contact the Digital Scholarship Operations Team or your Subject Librarian. UCIspace @ the Libraries is a service for the UCI community to publish, manage, and preserve diverse kinds of research output. Fully searchable and indexed by major search engines, with UCIspace @ the Libraries you may publish digital research materials for researchers to download, create description to help researchers find and understand your content, and rest assured that your files will be preserved and accessible long term." (site UCI Libraries)	Bailey 2013	http://ucispace.lib.uci.edu/page/about-ucispace (consulté le 22.12.2021)
Union List of Artist Names Online (ULAN)	Vocabulaire	The Getty Research Institute	1984		Description; Diffusion	"The Getty Vocabularies contain structured terminology for art, architecture, decorative arts, archival materials, visual surrogates, art conservation, and bibliographic materials. Compliant with international standards, they provide authoritative information for catalogers, researchers, and data providers. They contain coreferences to other resources where topics overlap; however, the Getty Vocabularies are unique in their global coverage of the defined domain, in citing published sources and contributors, in allowing interconnections among historical and current information, in accommodating the sometimes debated and ambiguous nature of art historical information, and in allowing complex relationships within and between Vocabularies." (site web) Voir aussi : Art & Architecture Thesaurus (AAT); Getty Thesaurus of	Gracy 2015. p. 270	https://www.getty.edu/research/tools/vocabularies/ulan/ (consulté le 20.12.2021)

Nom	Type	Développeur	Année début	Année fin	Fonction	Description	Référence(s) bibliographique(s)	Liens
						Geographic Names (TGN); Cultural Objects Name Authority (CONA); Iconography Authority (IA)"		
Using AI for Digital Records Selection in Government	Projet	The National Archives UK			Evaluation	"Digital transformation in government has brought an increase in the scale and variety of public records along with a reduced emphasis on organising and structuring data. Traditional processes designed for paper records cannot handle the volume, diversity, complexity and distributed nature of departmental digital records. The project by The National Archives explored the potential of Artificial Intelligence (AI) tools to assist with this challenge. Five AI vendors applied their tools to classify a dataset supplied by The National Archives. The tools and platforms evaluated were Adlib Elevate, Amazon Web Services, Microsoft Azure, InSight by Iron Mountain, and Records365 by RecordPoint. Promising results were obtained overall with no tool or approach consistently outperforming the others across all tasks." (site Using AI)	TNA 2021	https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/research-collaboration/using-ai-for-digital-selection-in-government/ (consulté le 20.12.2021)
Valeurs immatérielles transmises aux archives pour mémoire (Vitam)	Projet	Programme interministériel archives numérique (France)	2013		Evaluation; Accroissement; Classification; Description; Préservation; Diffusion	Trois ministères, Europe et Affaires étrangères (MEAE), Culture (MC), Armées (MinArm), responsables de la conservation des archives de l'État ont mutualisé leurs efforts pour répondre ensemble à l'enjeu de l'archivage numérique tout au long de leur cycle de vie (site vitam)	dominiquenaud 2019	http://www.programmevitam.fr/ (consulté le 23.12.2021)
VERS Initiative	Projet	Public Record Office Victoria	2017		Évaluation; Préservation	"The Victorian Electronic Records Strategy is about ensuring the creation, capture and preservation of authentic, complete and meaningful digital records by the Victorian public sector." (site VERS)	Rolan et al. 2019, p. 188	https://prov.vic.gov.au/recordkeeping-government/vers (consulté le 20.12.2021)

Annexe 3 : Liaisons Fonctionnalités-Tâches (tableau à double entrée)

	Fonctionnalités/Tâches	Evaluation					Versement			Acquisition						Classification						Description						Diffusion					Préservation numérique (OAIS)						
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
		Définir contexte de l'institution/personne qui a géré les archives	Élaborer des critères d'évaluation	Attribuer une valeur aux documents d'archives	Décider du degré de conservation	Définir règles (ou normes) de conservation	Dresser l'inventaire de l'ensemble des documents produits par l'organisme	Éliminer les archives non conservées	Verser les archives au service d'archives	Établir la liste des fonds souhaités et de leurs propriétaires	Établir la relation personnelle avec les donateurs visés	Négocier les acquisitions et Signer les ententes	Transférer et enregistrer les documents au service d'archives	Inventorier (récollement) les documents	Annoncer les acquisitions	Importer le plan de classification des archives courantes	Délimiter le fonds d'archives	Situer le fonds dans un plan général de classification	Répartir les documents en fonction de son ou ses créateurs	Décider d'un modèle de classification	Identifier les catégories d'utilisateurs et cibler leurs besoins d'information et leurs stratégies de recherche	Présenter des caractéristiques physiques des documents	Analyser le contenu des documents	Présenter le contexte de création et d'utilisation	Planifier, concevoir et réaliser l'instrument de recherche	Indexer	Former les usagers	Développer la clientèle	Mettre en valeur les documents	Mettre à disposition les documents	Reproduire les documents	Conversion des formats de fichiers ou des représentations des données	Gérer la hiérarchie du stockage	Remplacer les supports	Contrôler les erreurs	Fournir un plan de reprise	Gérer la configuration du système	Veille technologique
1	Ajout de métadonnées descriptives intellectuelles	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
2	Ajout de métadonnées descriptives techniques	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
3	Analyse de volumétrie	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
4	Analyse des sentiments (NLP)	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
5	Annotation de document	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
6	Application des délais de conservation	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	Attribution d'identifiants pérennes	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
8	Bloquage d'écriture	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0
9	Calcul de la valeur archivistique	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	Changement d'extension	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0
11	Classification et catégorisation de documents (NLP)	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
12	Comparaison de fichiers texte	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	Compilation de métadonnées (en format XML)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
14	Contrôle d'autorité	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
15	Contrôle de conformité du SIP	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0
16	Contrôle de signature électronique	0	0	0	0	0	0	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	Contrôle des droits d'accès	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
18	Contrôle du vocabulaire	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
19	Conversion de format	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0
20	Craquer mot de passe	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	Création d'arborescence virtuelle	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
22	Création d'un bordeau d'élimination	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	Création d'un bordeau de versement	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	Création d'un calendrier de conservation	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	Création d'un plan de classement	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
26	Création d'un rapport de recherche	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
27	Création d'un rapport de récolement	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
28	Création d'un rapport synthétique	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	Création de copies de téléchargement	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0
30	Création de copies d'accès	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	0
31	Décryptage de fichiers	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	Dédoublonnage	0	0	0	1	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
33	Evaluation des risques archivistiques	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
34	Exportation vers un SAE	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	Extraction d'informations contextuelles (NLP)	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
36	Extraction de métadonnées	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	Extraction de texte	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
38	Garantie de la valeur probante (empreinte digitale numérique)	0	0	1	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
39	Gestion d'ontologies	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0	0	0
40	Identification de format	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0
41	Identification de langue	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
42	Importation d'un plan de classement préexistant	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	Importation de documents	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	Journalisation	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0
45	Migration de support	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
46	Modèle de description standardisé	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
47	Moissonnage du web	0	0	0	1	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
48	Prévisualisation de fichiers	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
49	Recherche à facettes	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0
50	Recherche générale	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0
51	Recherche géospatiale	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0
52	Recherche plein-texte	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0
53	Reconnaissance d'entités nommées (NLP)	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
54	Reconnaissance d'images	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
55	Reconnaissance optique de caractères imprimés	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
56	Reconnaissance optique de caractères manuscrits	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
57	Récupération de données	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
58	Remaniement de l'arborescence	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
59	Renommage de fichiers	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
60	Sensitivity Review (NLP)	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
61	Topic modeling (NLP)	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
62	Utilisation de Linked Data	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0
63	Validation de format	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
64	Visualisation de chronologie	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
65	Visualisation de l'arborescence	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
66	Visualisation par cartes géographiques	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0

Annexe 4 : Diagramme



Légendes des tâches et des fonctionnalités

Evaluation	1	Définir contexte de l'institution/personne qui a géré les archives
	2	Élaborer des critères d'évaluation
	3	Attribuer une valeur aux documents d'archives
	4	Décider du degré de conservation
	5	Définir règles (ou normes) de conservation
Versement	6	Dresser l'inventaire de l'ensemble des documents produits par l'organisme
	7	Éliminer les archives non conservées
	8	Verser les archives au service d'archives
Acquisition	9	Établir la liste des fonds souhaités et de leurs propriétaires
	10	Établir la relation personnelle avec les donateurs visés
	11	Négocier les acquisitions et Signer les ententes
	12	Transférer et enregistrer les documents au service d'archives
	13	Inventorier (récollement) les documents
	14	Annoncer les acquisitions
Classification	15	Importer le plan de classification des archives courantes
	16	Délimiter le fonds d'archives
	17	Situer le fonds dans un plan général de classification
	18	Répartir les documents en fonction de son ou ses créateurs
	19	Décider d'un modèle de classification
Description	20	Identifier les catégories d'utilisateurs et cibler leurs besoins d'information et leurs stratégies de recherche
	21	Présenter des caractéristiques physiques des documents
	22	Analyser le contenu des documents
	23	Présenter le contexte de création et d'utilisation
	24	Planifier, concevoir et réaliser l'instrument de recherche
	25	Indexer
Diffusion	26	Former les usagers
	27	Développer la clientèle
	28	Mettre en valeur les documents
	29	Mettre à disposition les documents
	30	Reproduire les documents
Préservation numérique (OAIS)	31	Conversion des formats de fichiers ou des représentations des données
	32	Gérer la hiérarchie du stockage
	33	Remplacer les supports
	34	Contrôler les erreurs
	35	Fournir un plan de reprise
	36	Gérer la configuration du système
	37	Veille technologique

Fonctionnalités	1	Ajout de métadonnées descriptives intellectuelles
	2	Ajout de métadonnées descriptives techniques
	3	Analyse de volumétrie
	4	Analyse des sentiments (NLP)
	5	Annotation de document
	6	Application des délais de conservation
	7	Attribution d'identifiants pérennes
	8	Blocage d'écriture
	9	Calcul de la valeur archivistique
	10	Changement d'extension
	11	Classification et catégorisation de documents (NLP)
	12	Comparaison de fichiers texte
	13	Compilation de métadonnées (en format XML)
	14	Contrôle d'autorité
	15	Contrôle de conformité du SIP
	16	Contrôle de signature électronique
	17	Contrôle des droits d'accès
	18	Contrôle du vocabulaire
	19	Conversion de format
	20	Craquer mot de passe
	21	Création d'arborescence virtuelle
	22	Création d'un bordeau d'élimination
	23	Création d'un bordeau de versement
	24	Création d'un calendrier de conservation
	25	Création d'un plan de classement
	26	Création d'un rapport de recherche
	27	Création d'un rapport de récolement
	28	Création d'un rapport synthétique
	29	Création de copies de téléchargement
	30	Création de copies d'accès
	31	Décryptage de fichiers
	32	Dédoublonnage
	33	Evaluation des risques archivistiques
	34	Exportation vers un SAE
	35	Extraction d'informations contextuelles (NLP)
	36	Extraction de métadonnées
	37	Extraction de texte
	38	Garantie de la valeur probante (empreinte digitale numérique)
	39	Gestion d'ontologies
	40	Identification de format
	41	Identification de langue
	42	Importation d'un plan de classement préexistant
	43	Importation de documents
	44	Journalisation
	45	Migration de support
	46	Modèle de description standardisé
	47	Moissonnage du web
	48	Prévisualisation de fichiers
	49	Recherche à facettes
	50	Recherche générale
	51	Recherche géospatiale
	52	Recherche plein-texte
	53	Reconnaissance d'entités nommées (NLP)
	54	Reconnaissance d'images
	55	Reconnaissance optique de caractères imprimés
	56	Reconnaissance optique de caractères manuscrits
	57	Récupération de données
	58	Remaniement de l'arborescence
	59	Renommage de fichiers
	60	Sensitivity Review (NLP)
	61	Topic modeling (NLP)
	62	Utilisation de Linked Data
	63	Validation de format
	64	Visualisation de chronologie
	65	Visualisation de l'arborescence
	66	Visualisation par cartes géographiques