h e g

Haute école de gestion
Genève

État des lieux des pratiques d'évaluation aux Archives de l'État de Neuchâtel et exploration de leur traduction en fonctionnalités pour le logiciel *ArchiSelect*

Travail de master réalisé par :

Sébastien BISCHOFF

Sous la direction de :

Basma MAKHLOUF SHABOU, professeure HES

Carouge, 15.08.2022

Science de l'information

Haute École de Gestion de Genève (HEG-GE)

Hes·so Genève
Haute Ecole Spécialisée de Suisse occidentale

Remerciements

Mes remerciements vont d'abord à Basma Makhlouf Shabou, directrice de ce Mémoire, pour son suivi et ses conseils avisés, ainsi qu'à Monsieur Carol Couture qui m'a fait l'honneur d'accepter la charge d'expert.

Je remercie également mon mandant, Lionel Bartolini, archiviste cantonal de Neuchâtel, ainsi que toute l'équipe des Archives de l'État de Neuchâtel d'avoir pris le temps de répondre à mes nombreuses questions et pour les discussions stimulantes et enrichissantes menées.

Ma reconnaissance va également aux personnes des archives cantonales romandes qui ont accepté mes entretiens ainsi qu'à Arnaud Gaudinat de la Haute école de gestion de Genève.

Mes remerciements vont aussi à mes camarades de classe qui furent une réelle source de motivation. Remerciements spéciaux à Denis Bussard pour son rôle de *sparring partner* dans nos discussions autour l'évaluation archivistique.

Enfin, mes remerciements vont à ma famille et ami-e-s pour avoir supportés mes états d'âme et offert des moments de coupure bienvenus.

Résumé

L'évaluation archivistique à l'ère du numérique va devoir permettre de maîtriser une production informationnelle grandissante et souvent peu structurée. Pour répondre à ce défi, les Archives de l'État de Neuchâtel (AEN), mandant du présent mémoire, développent une série d'outils accompagnant l'archiviste dans les différentes étapes de l'archivage numérique.

L'outil dédié à l'évaluation archivistique, *ArchiSelect,* est en phase de conceptualisation et a déjà fait l'objet d'un mandat de recherche confié aux Prof. Basma Makhlouf Shabou et Prof. Arnaud Gaudinat de la Haute école de gestion de Genève. Le résultat de cette recherche propose un cadre conceptuel définissant des critères d'évaluation et une preuve de concept de l'opérationnalisation des critères avec le recours au *text mining*. Intégrant et discutant ces résultats, ce mémoire propose une réflexion approfondie sur les fonctionnalités que pourrait proposer *ArchiSelect* pour soutenir l'archiviste dans sa tâche de l'évaluation.

Pour proposer des fonctionnalités pertinentes, un état des lieux approfondi des projets et outils existants a été mené, révélant l'absence de solution pratique et théorique faisant l'unanimité en termes d'évaluation archivistique de documents numériques. Le présent mémoire s'appuie donc sur l'analyse des pratiques d'évaluation archivistique de documents analogiques ayant cours aux AEN, analysant et séquençant le *workflow* d'évaluation puis traduisant les différentes tâches dans le domaine numérique sous forme de fonctionnalités.

L'analyse des pratiques d'évaluation et du *workflow* a démontré qu'*ArchiSelect* allait devoir intégrer des fonctionnalités pour faciliter l'évaluation notamment de la dimension de *représentativité* des documents, difficilement automatisable du fait de ou étant donné ou compte tenu de son aspect subjectif. Partant de ce constat, ce travail propose un *workflow* pour l'évaluation de documents numériques compatible avec les pratiques d'évaluation neuchâteloises, associé à un set de fonctionnalités-clés qui permet la réalisation de le la fonction d'évaluation, intégrant notamment des fonctionnalités à base *machine learning* et *Natural language processing*.

Mots-clés : évaluation archivistique ; archivage numérique ; critères d'évaluation ; collections de fichiers non structurés ; *text-mining*

Table des matières

Liste des tableaux

Liste des figures

Abréviations

AAS	Association des archivistes suisses
AEN	Archives de l'État de Neuchâtel
AENeas	Archives de l'État de Neuchâtel *electronic archiving system*
AFS	Archives fédérales suisses
AI	*Arificial intelligence*
AIP	*Archival Information Package*
BPMN	*Business Process Model and Notation*
CECO	Centre de coordination pour l'archivage à long terme de documents électroniques = *Koordinationsstelle für die dauerhafte Archivierung elektronische Unterlagen* (KOST)
COPTR	*Community Owned digital Preservation Tool Registry*
DESC	Département de l'économie, de la sécurité et de la culture (Canton de Neuchâtel)
DEV	Dimension d'évaluation
DIP	*Dissemination Information Package*
DLA	*Document Layout Analysis*
DU	Durée d'utilité
DUA	Durée d'utilité administrative
DUL	Durée d'utilité légale
FS	*File System*
GED	Gestion électronique des documents
GEVER	*Geschäftsverwaltung*
HEG-GE	Haute école de gestion de Genève
ISO	*International Standard Organization* – Organisation internationale de normalisation
LArch	Loi sur l'archivage[1]
LDA	*Latent Dirichlet Allocation*

[1] Lorsque rien d'autre n'est précisé, LArch sera utilisé pour désigner la Loi sur l'archivage du 22 février 2011 (LArch ; 442.20) du canton de Neuchâtel.

MAS ALIS	*Master of Advanced Studies in Archival, Library and Information Science*
ML	*Machine learning*
MPLP	*More Product Less Process*
NARA	*National Archives and Records Administration* (États-Unis)
NER	*Named Entity Recognition*
OAIS	*Open archival information system* = Système ouvert d'archivage d'information (SOAI)
OCC	One-class classification
OCR	*Optical Character Recognition*
OORG	Office de l'organisation (canton de Neuchâtel)
PA	Plan d'archivage
QADEPs	Qualités des documents et des données électroniques publics
RLArch	Règlement d'exécution de la loi sur l'archivage (LArch)
RM	*Records Management*
ROC	Reconnaissance optique de caractères
SCI	Système de contrôle interne
SD	Système documentaire
s.d.	Sans date
SI	Système d'information
SIEN	Service informatique de l'entité neuchâteloise
SIP	*Submission Information Package*
SyFi	Système de fichiers (*file system*)
TNA	*The National Archives* (Royaume Uni)

Glossaire

AENeas « Le concept AENeas est la réponse neuchâteloise au défi de l'archivage numérique. Il propose notamment plusieurs outils (la "SuiteArchi") qui accompagnent le cycle de vie documentaire. » (Oguey, Schneiter 2018)

ArchiClass « ArchiClass est le premier outil issu du concept AENeas. 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. » (Archives de l'État de Neuchâtel s.d. a)

ArchiPeren Outil de la *SuiteArchi* qui servira de « dépôt numérique et qui assumera également le conditionnement numérique des archives à verser. » (Oguey, Schneiter 2018)

ArchiRef « ArchiRef est l'outil de contrôle de la bonne application des règles définies dans le plan d'archivage dans l'environnement de gestion documentaire. » (Archives de l'État de Neuchâtel s.d. a)

ArchiSelect Outil de la *SuiteArchi* qui doit « servir à accompagner l'archiviste dans l'évaluation des dossiers qui lui sont proposés pour versement et/ou élimination » (Oguey, Schneiter 2018)

ArchiVision Outil de la *SuiteArchi* qui « permettra de suivre la supervision des entités soumises à la loi sur l'archivage » (Oguey, Schneiter 2018)

Bordereau de versement et d'élimination « Les bordereaux de versement et d'élimination décrivent sommairement les documents concernés et précisent leurs dates extrêmes. Ils signalent les documents contenant des données personnelles ou sensibles ainsi que les documents librement accessibles au public selon la convention intercantonale relative à la protection des données et à la transparence dans les cantons de Neuchâtel et du Jura (CPDT-JUNE), des 8 et 9 mai 2012. Ils sont signés par la direction de l'autorité. » (Archives de l'État de Neuchâtel 2016, p. 30).

Broad appraisal [« évaluation globale »] En opposition à *in-depth appraisal* « *broad appraisal, i.e., deleting duplicates, empty or junk files/folders.* » (Belovari 2017, p. 61)

Cadre conceptuel « Brève explication fondée sur l'agencement logique d'un ensemble de concepts et de sous-concepts liés entre eux et réunis en raison de leur affinité avec le problème de recherche » (Fortin 2016, p. 497).

Checksum [« somme de contrôle »] « *unique alphanumeric value that represents the bitstream of an individual computer file or set of files* »; « *The term checksum is often used interchangeably with other types of fixity tools such as cryptographic hash values generated by algorithms such as MD5, SHA-1, and SHA-256. Checksums or hashes are used in the context of digital*

preservation, such as during the process of transferring or storing files, in order to determine whether files have been altered. For example, an archivist can compare checksums or hashes generated before and after file transfer to determine whether the file has maintained the same value through the transfer process. » (Dictionary of Archives Terminology 2022)

Clôture La clôture d'un dossier d'activité intervient lorsque la délivrance d'une prestation est considérée comme terminée et qu'a priori, le dossier y relatif ne sera plus augmenté ni modifié.

***Clustering* [ou *Cluster analysis*]** [« partitionnement de données »] « *Cluster analysis is the formal study of methods and algorithms for grouping, or clustering, objects according to measured or perceived intrinsic characteristics or similarity. Cluster analysis does not use category labels that tag objects with prior identifiers, i.e., class labels. The absence of category information distinguishes data clustering (unsupervised learning) from classification or discriminant analysis (supervised learning). The aim of clustering is to find structure in data and is therefore exploratory in nature.* » (Jain 2010, p. 651)

Deep learning [« apprentissage profond »] « Mode d'apprentissage automatique généralement effectué par un réseau de neurones artificiels composé de plusieurs couches de neurones hiérarchisées selon le degré de complexité des concepts, et qui, en interagissant entre elles, permettent à un agent d'apprendre progressivement et efficacement à partir de mégadonnées [*big data*]. » (Grand dictionnaire terminologique 2012)

Digital curation « *Digital curation is the management and preservation of digital data/information over the long-term.* » (Digital Curation Center 2022)

Dimension d'évaluation (DEV) Les Dimensions d'évaluation sont des « Concepts généraux qui définissent les perspectives selon lesquelles la et les valeurs des archives seront identifiées » (Makhlouf Shabou, Tièche 2018a, p. 24). « Les dimensions d'évaluation amènent à l'émergence de variables, véritables éléments de mesures qui permettent de qualifier leurs valeurs associées. » (Makhlouf Shabou, Tièche 2018a, p. 8). Les trois dimensions principales étant la *valeur probante*, l'*exploitabilité* et la *représentativité*.

Diplomatique La diplomatique est une science médiévale qui qui consiste à étudier les « attributs exclusif » des documents. Elle se concentre sur l'analyse de leur nature mais également leur contexte général et leur création ainsi que leurs conditions de transmission. La diplomatique a reconnu un regain d'intérêt dans l'application de ses méthodes aux documents numériques. (Duranti 1989; 1998; 2003; Chabin 2008; Makhlouf Shabou 2015a).

Disk image [« image disque »] « *A disk image, in computing, is a computer file containing the contents and structure of a disk volume or of an entire data storage device, such as a hard disk drive, tape drive, floppy disk, optical disc, or USB flash drive. A disk image is usually made by creating a sector-by-sector copy of the source medium, thereby perfectly replicating the structure*

and contents of a storage device independent of the file system. » (*Disk image* 2022)

Document d'activité « Informations créées, reçues et préservées comme preuve et actif par une personne physique ou morale dans l'exercice de ses obligations légales ou la conduite des opérations liées à son activité » (ISO 15489-2 :2016, p. 3)

eDiscovery « *The discovery or disclosure of electronic information for the purposes of litigation* » (The National Archives UK 2016, p. 3)

Entité organisationnelle Entité soumise au champ d'application de la LArch : les autorités cantonales et les autorités communales. (Archives de l'État de Neuchâtel 2016, p. 28) Parfois appelées uniquement « entités » dans ce travail.

Gestion électronique des documents (GED) « Application permettant d'organiser et de gérer les informations et documents électroniques d'une organisation » (Béchard, Fuentes Hashimoto, Vasseur 2020, p. 79)

GEVER De l'allemand *GEschäftsVERwaltung*, gestion des affaires. Norme fédérale sur les conditions juridiques et techniques nécessaires à une gestion documentaire électronique qui soit conforme à la norme ISO 15489 (Debenath et al. 2016)

Linked data [« données liées »] « Ensemble de données munies de leurs métadonnées qui, reliées les unes aux autres, constituent une base de données à l'échelle du Web. » (Grand dictionnaire terminologique 2012)

***Machine learning* (ML)** [« Apprentissage automatique »] « *the study of computer algorithms that improve automatically through experience* » (Mitchell 1997); « Mode d'apprentissage par lequel un agent évalue et améliore ses performances et son efficacité sans que son programme soit modifié, en acquérant de nouvelles connaissances et aptitudes à partir de données et/ou en réorganisant celles qu'il possède déjà. » (Grand dictionnaire terminologique 2012)

Mapping [« Mise en correspondance »] « Association des données appartenant à un ensemble (modèle logique de données, base de données de production, champ source) avec les données appartenant à un autre ensemble (modèle physique de données, entrepôt de données, champ cible), de manière que les données du premier ensemble puissent se substituer à celle du second ensemble, ou encore que l'on puisse passer harmonieusement des premières aux secondes. » (Grand dictionnaire terminologique 2012)

Métrique Mesure qui permet d'évaluer qualitativement et/ou quantitativement les documents (Gaudinat, Knafou 2018a, p. 4)

Métrique archivistique Métriques provenant du modèle de métriques d'évaluation, structuré autour des dimensions d'évaluations de *valeur probante*, *exploitabilité* et *représentativité*.

Métrique de fouille de données Métadonnées ou combinaisons de métadonnées qui permettent d'évaluer les documents ou les lots de données. (Gaudinat, Knafou 2018a, p. 9)

***Named Entity Recognition* (NER)** [Reconnaissance d'entités nommées] « *Named entity recognition and classification (NER for short) corresponds to the identification of entities of interest in texts, generally of the types Person, Organisation and Location. Such entities act as referential anchors which underlie the semantics of texts and guide their interpretation.* » (Ehrmann et al. 2021, p. 2)

***Natural Language Processing* (NLP)** Le NLP est une technologie, une forme d'intelligence artificielle qui cherche à apprendre aux machines de comprendre ou générer du langage naturel (soit le langage humain, en opposition au langage « informatique » ou « de programmation ». Tamas E. Doszkocs nous en donne la définition suivante : « *Loosely defined, natural language processing (NLP) encompasses all computer-based approaches to handling unrestricted written or spoken language, from purely "mechanistic" procedures, as employed by text editors, word processors, and in automatic-indexing approaches in information retrieval (IR) to "intelligent" analysis, understanding and expression of "meaning" as exemplified in natural language understanding, question answering and expert systems (AI).* » (1986, p. 191).

***Open archival information system* (OAIS)** Modèle conceptuel d'archivage à long terme. Il s'agit d'une organisation constituée d'une équipe de systèmes, dont la responsabilité est de pérenniser l'information et de la rendre accessible à un groupe d'utilisateur désigné (ISO 14721 :2012, p. 1-1)

Open source « *A computer program in which the source code is available to the general public for use and/or modification from its original design free of charge (open).* »; « *A method and philosophy for software licensing and distribution designed to encourage use and improvement of software written by volunteers by ensuring that anyone can copy the source code and modify it freely.* » (InterPARES Trust 2018)

Plan d'archivage (PA) Référentiel archivistique se composant de : a) un cadre de classement des documents, établi selon les directives de l'office et reflétant les missions et les activités de l'autorité ; b) les prescriptions relatives à l'accessibilité et à la confidentialité des documents ; c) les délais d'utilité administrative et légale des documents, déterminés par l'autorité, en concertation avec l'office (RLArch, art. 5, al. 2.) ; d) la valeur archivistique des documents, évaluée par l'office, en concertation avec l'autorité

Préposé à la gestion des documents Personne chargée de la gestion des documents, de « la rédaction et de la tenue à jour du plan d'archivage, ainsi que de l'organisation des éliminations et des versements » (RLarch, art. 6). Nous ferons parfois uniquement référence à « préposé ».

Records management Le *Records management* est le champ de l'organisation et de la gestion en charge d'un contrôle efficace et systématique de la création, réception, conservation, l'utilisation et du sort final des documents d'activité, des processus de capture et de préservation de la preuve et de l'information liées aux activités et aux opérations sous la forme de documents d'activités. (ISO 15489-2 :2016)

Registratur « Système permettant de contrôler la création et le maintien en état des dossiers courants au moyen de registres, répertoires, index et de plans de rangement » (Walne 1988, p. 134)

Sensitvity review « *The process of identifying sensitive content in digital records that should be exempt from release.* » (The National Archives UK 2016, p. 3)

Sort final pressenti « Sort final des dossiers défini lors de l'évaluation prospective à valider par l'entité d'archives lors de la proposition » (Gaudinat, Knafou 2018a, p. 5) ; ou de manière plus générique : « le sort final pressenti est la destination (élimination, conservation ou échantillonnage) de documents, d'un ensemble de documents (ou données ou dossier) à l'expiration de leur délai d'utilité administrative et légale, inscrit dans le calendrier de conservation mais qui peut être soumise à une réévaluation au moment de sa mise en œuvre. » (Dunant Gonzenbach 2022)

SuiteArchi Suite d'outils développés par les AEN « qui accompagnent le cycle de vie documentaire et le processus d'archivage, de la création d'un plan d'archivage à la mise à disposition des archives pour la consultation. » (Oguey, Schneiter 2018) Font partie de la *SuiteArchi* : *ArchiClass, ArchiRef, ArchiSelect, ArchiVision, ArchiPeren* et *ArchiInfo*

Système de fichiers [*file system*] « méthode utilisée pour stocker de l'information sur un support de stockage » (Béchard, Fuentes Hashimoto, Vasseur 2020, p. 81)

Système d'information (SI) « Ensemble des ressources (personnels, matériels, logiciels) organisés pour collecter, stocker, traiter et communiquer les informations. Le système d'information coordonne grâce à l'information les activités de l'organisation et lui permet ainsi d'atteindre ses objectifs. » (Béchard, Fuentes Hashimoto, Vasseur 2020, p. 81)

Text mining « *A process to analyze natural-language text, typically unstructured, for a variety of purposes, including summarization, document clustering, text categorization, language identification, authorship attribution, and extracting data elements (names, dates, abbreviations, acronyms)* » (InterPARES Trust, 2018)

Topic-modelling « […] *a suite of algorithms that aim to discover and annotate large archives of documents with thematic information. Topic modeling algorithms are statistical methods that analyze the words of the original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over time.* » (Blei 2012, p. 77)

Überlieferungsbildung « On peut traduire [*Überlieferungsbildung*] par « constitution du patrimoine archivistique » Cette notion est certes plus large que celle de d'évaluation, puisqu'elle inclut toutes les mesures qui permettent d'assurer que les documents de valeur durable parviendront bien aux archives. L'évaluation demeure cependant au cœur de l'*Überlieferungsbildung*. » (Burgy, Egli, Schmutz 2007, p. 282)

Variable [archives] « Les dimensions d'évaluation amènent à l'émergence de variables, véritables éléments de mesures qui permettent de qualifier leurs valeurs associées. » (Makhlouf Shabou, Tièche 2018a, p. 8).

Vrac[2] « Fonds qui se présente sans aucun classement et dont la provenance peut ne pas être identifiée. » (Association des archivistes français 2012).

Workflow [« flux de travaux »] Processus « au cours duquel des tâches, des procédures et des informations sont traitées ou exécutées successivement, selon des règles prédéfinies, en vue de réaliser un produit ou de fournir un service. » (Grand dictionnaire terminologique 2012)

Write blocker[bloqueur d'écriture] Dispositif qui empêche les objets numériques d'être modifiés ou altérés pendant le processus de transfert d'un dispositif de stockage à un autre. Les bloqueurs d'écriture ont été développés dans le cadre de la criminalistique numérique (*digital forensics*), mais peut être utilisé à des fins de préservation numérique. (Community Archives and Heritage Group 2018)

[2] Également voir 1.3.5 « Collections de fichiers non structurés et vracs numériques ».

1. Introduction

Dans ce travail nous explorons les possibilités de fonctionnalités pouvant soutenir l'archiviste dans son travail d'évaluation archivistique. Le projet d'archivage numérique des Archives de l'État de Neuchâtel (AEN), *AENeas*, prévoit une série d'outils accompagnant le processus d'archivage numérique. Encore en phase de conceptualisation, l'outil *ArchiSelect,* qui doit « servir à accompagner l'archiviste dans l'évaluation des dossiers qui lui sont proposés pour versement et/ou élimination » (Oguey, Schneiter 2018) a fait l'objet d'un mandat de recherche confié aux Prof. Basma Makhlouf Shabou et Prof. Arnaud Gaudinat de la Haute école de gestion de Genève : *Modèle d'évaluation systématique des documents : Concept & Preuve de concept*[3]. Les résultats du mandat seront pleinement intégrés et discutés dans le présent travail. Le but de ce mémoire est de proposer une réflexion plus spécifique autour des *fonctionnalités* que pourrait proposer cet outil en prévision de la rédaction d'un cahier des charges.

Cette réflexion partira d'une analyse approfondie des pratiques d'évaluation archivistique actuelles aux AEN et se nourrira de la littérature professionnelle et des autre projets existants. Ainsi, nous commencerons par exposer dans la problématique (1.1) les principaux défis auxquelles sont et seront confrontés à l'avenir les archivistes dans la gestion d'une production documentaire dont la croissance semble quasi exponentielle (Weill 1990). Nous définirons ensuite les objectifs et notre question de recherche (1.2) ainsi que les principaux concepts (1.3). Après avoir explicité notre méthodologie (2), nous allons procéder à un état des lieux approfondi de la littérature et projets existant autour de l'évaluation des archives numériques (3) et présenterons le contexte documentaire du canton de Neuchâtel (4).

Enfin, nous exposerons les résultats et notre analyse, soit un état des lieux de l'évaluation archivistique aux AEN (5.1) et nous ferons une proposition d'un *workflow* pour l'évaluation de documents numériques (5.2). En nous basant sur ces éléments, nous développerons une réflexion autour des fonctionnalités soutenant les tâches de l'évaluation de documents numériques (5.3). Nous discuterons les principaux résultats et finiront par une conclusion (7).

1.1 Problématique

ArchiSelect est l'outil qui, dans le cadre du projet *AENeas*, servira de soutien à l'évaluation à l'archiviste dans le domaine numérique. « Fonction pivot » de l'archivistique (Couture, Lajeunesse 2014), elle est régulièrement considérée comme la plus essentielle et la plus noble tâche de l'archiviste (Couture 1996, p. 3). Dans le domaine numérique l'évaluation garde toute son importance et devra s'adapter à des réalités nouvelles, impliquant des défis supplémentaires.

L'évaluation archivistique devra maîtriser le changement d'échelle de la volumétrie d'abord, qui devient une composante fondamentale dans la réflexion archivistique (Coutaz 2016). Les services d'archives doivent ainsi se munir d'outils capables de soutenir l'archiviste dans l'évaluation afin de maîtriser la production documentaire et venir à bout du passif déposé sur tout type de systèmes de stockages. Les exemples sont hélas multiples de « passifs abyssaux » (Coutaz 2014) non traités, la *National Archives and Records Administration*

[3] Présenté dans la partie 3.2.1 de ce travail, les résultats de cette recherche seront en général désignés par « mandat réalisé par la HEG-GE ».

(NARA) estimait en 2013 que 95 % de ses archives numériques n'étaient pas traitées (Trace 2021).

Certes, l'évolution des espaces de stockage suit une courbe similaire et l'on pourrait être tenté de *tout* garder et stocker[4], ce serait pourtant répudier le principe même de l'archivage : l'évaluation reste valide est incontournable. En effet, nous sommes persuadés que la création d'une mémoire collective passe paradoxalement aussi par l'oubli. À l'image du personnage de Funes dans la nouvelle *Funes ou la mémoire* de Jorge Luis Borges qui, suite à un accident, est doté d'une mémoire infaillible, dont le narrateur dit : « Je soupçonne cependant qu'il n'était pas très capable de penser. Penser c'est oublier des différences, c'est généraliser, abstraire. Dans le monde surchargé de Funes il n'y avait que des détails, presque immédiats. » (1956, p. 118). Sylvain Senécal cite également ce texte et suggère qu'une réplication, un miroir exact du présent rend impossible toute histoire (2013, p. 209). Marie-Anne Chabin propose elle le néologisme de « surmnésie » pour mettre en garde contre « l'obésité archivale » (2007).

Autre défi de taille, intimement lié à la question de la volumétrie, celui de la structure, ou plutôt de son absence et donc de pouvoir *appréhender* l'information déposée sur des systèmes de stockages non structurés – « *Even just "knowing" what is contained in a digital collection can be challenging.* » (Belovari 2017, p. 56). À titre d'exemple, Caroline Pegden nous apprend que deux tiers de l'information de l'administration du Royaume-Uni est conservée sur des lecteurs partagés non structurés, certains ministères stockant jusqu'à 190 Téraoctets d'information sur des serveurs de courrier électronique (2016).

Le classement de documents analogiques se heurte à des limites physiques lorsque le nombre de documents augmente, alors que les systèmes de fichiers ne limitent ni le nombre de documents, ni le nombre de dossiers sur le même niveau. Ainsi, la facilité de créer de nouveaux dossiers invite généralement à le faire spontanément plutôt que d'évaluer dans quel dossier existant devrait être classé un document. Cela a comme conséquence que les dossiers d'affaires déposés sur des systèmes de fichiers ont tendance à être particulièrement volumineux et peu structurés (Schludi 2013, p. 24). Marie-Anne Chabin suggère que la paperasse a été remplacée par « l'électronasse », le terme, à défaut d'être élégant, est éloquent. La paperasse avait au moins le mérite d'être « compartimentée et localisée par producteur, ce qui limite objectivement le désordre » (2013) l'auteure pousse encore plus loin sa réflexion :

> *« […] là où naguère, du fait de la contrainte des supports et des distances, on produisait une note ou un courrier bien pensé, condensant sur un format A4 tous les éléments d'identification de l'auteur, du destinataire, de discours contextualisé et engageant, de date du discours, de validation […] d'énoncé, on va aujourd'hui, sous la pression des nouveaux outils et des nouvelles formes d'organisation et de travail collaboratif, produire l'équivalent (il y a toujours à un moment ou un autre une décision ou une affirmation engageante) sous la forme d'une succession de messages électroniques embarquant des liens hypertextes, des pièces jointes, des enregistrements de date, de signature, de données d'identification, de paramètres de sécurité, etc. »* [5] *(2013)*

[4] À ce propos, voir Gaudinat (2016).

[5] Dans un même ordre d'idée d'une progressive dilution de l'information, Chabin (2013) cite Luciana Duranti « *The single document of the Middle Ages might be the dossier of modern times.* » (1989)

Stephan Lenartz identifie ainsi quatre caractéristiques que l'on retrouve généralement dans les collections de fichiers non structurés : un nombre élevé de fichiers individuels qui en rend difficile la compréhension ; une juxtaposition de fichiers avec et sans valeur archivistique[6] ; une absence de structure de classement standardisée (fréquemment une organisation thématique) ; et une complexité due à l'hétérogénéité de formats dont certains sont inconnus (2020, p. 17).

Il n'existe pourtant toujours pas de solution théorique et technique normalisée pour maîtriser ces supports de stockages non structurés. Les stratégies existantes dans le traitement de vracs manquent quasi toutes d'adaptabilité aux réalités très différentes des propositions à analyser (Lenartz 2020, p. 18). Ainsi, les collections de fichiers ou vracs numériques restent par leur taille et leur difficulté à être appréhendés le « cauchemar des archivistes contemporains » (Taylor 2016, p. 7).

1.2 Objectifs et question de recherche

Nous tenterons ainsi d'explorer les possibilités offertes à l'archiviste pour réaliser l'évaluation archivistique de documents numériques, en adéquation avec les besoins spécifiques du contexte neuchâtelois, nos objectifs principaux sont les suivants :

- restituer les pratiques d'évaluation archivistique analogique actuelles aux Archives de l'État de Neuchâtel, avec prélèvement du *workflow* en place ;
- proposer des recommandations pour une optimisation dudit *worfklow* de l'évaluation archivistique analogique ainsi qu'une proposition d'un possible *workflow* pour l'évaluation de documents numériques ;
- recommander des fonctionnalités venant soutenir les tâches définies dans le *workflow* d'évaluation de documents numériques.

Notre question de recherche est la suivante : Dans le contexte neuchâtelois particulier, quelles fonctionnalités seraient utiles à l'archiviste dans sa tâche de l'évaluation archivistique de documents numériques ?

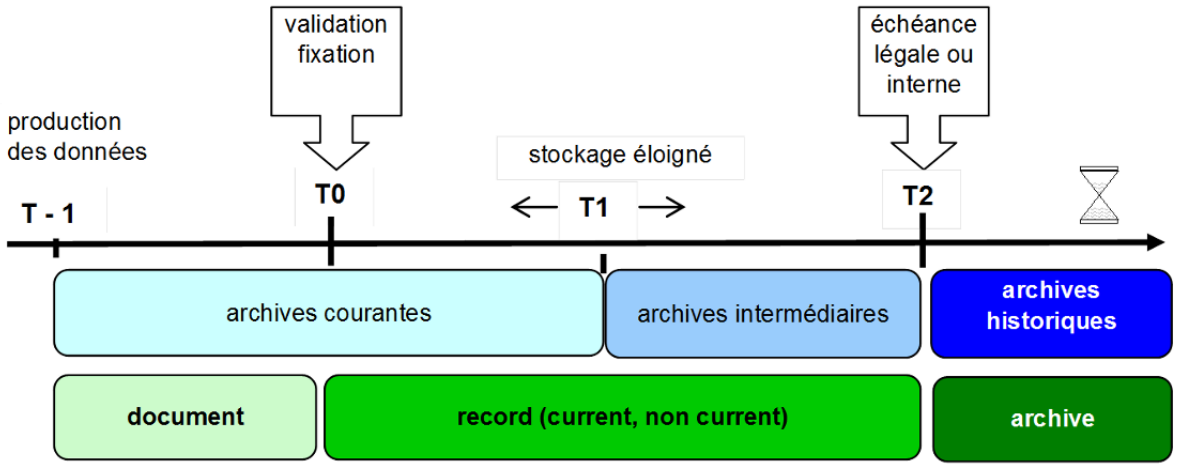
1.3 Définition des concepts

1.3.1 Cycle de vie documentaire

Inspiré par les réflexions de l'administration fédérale des États-Unis sous la présidence Truman (commission Hoover) en 1947, qui recommande alors une gestion des documents publics selon un cycle de vie, Yves Pérotin (1961), archiviste français, incite les archivistes à ajuster le traitement des documents en considérant « trois âges », chacun nécessitant des traitements spécifiques adaptés (Makhlouf Shabou 2012a, p. 29). Dans sa « théorie des trois âges », Pérotin évoque le principe du cycle de vie documentaire et subdivise l'existence des archives en trois périodes distinctes : *période d'activité*, où les archives sont dites courantes, soit couramment utilisées pour les raisons pour lesquels elles ont été créées ; une *période intermédiaire*, où les archives sont considérées comme semi-actives car plus qu'utilisées de manière occasionnelle ; et *période définitive* où les archives acquièrent une valeur de témoignage (Makhlouf Shabou 2012a, p. 30).

[6] « […] *Archivwürdig* […] *und nicht-archivwürdige Dateien.* » (Lenartz 2020, p. 17)

Figure 1 : Cycle de vie du document



(Chabin 2006)

La Figure 1 représente le cycle de vie du document avec l'approche « latine » (Chabin, Watel 2006) en bleu et l'approche plus anglo-saxonne issue du *records management* en vert.

Notamment l'avènement des documents numériques soulève des critiques mettant en cause la théorie des trois âges, le considérant comme obsolète (Lemay, Klein 2014; Dekens 2011). Les archivistes australiens, dont Frank Upward, proposent un modèle moins linéaire, le *Records continuum* qui place les archives dans une continuité de dimensions interconnectées (Upward 1996; McKemmish, Upward, Reed 2010; Upward et al. 2018; Gilliland 2014).

1.3.2 Évaluation archivistique

Il semble acquis que l'évaluation est l'une des sept fonctions archivistiques[7] (Couture, Rousseau 1994, p. 47-49). Carol Couture en donne une définition « consensuelle » (Doom 2006, p. 6) :

> *« […] 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)*

Il n'en demeure pas moins que la notion est mal circonscrite et pose notamment des problèmes terminologiques, spécialement en français où on l'identifie (du moins en France et dans certains cas au Québec) à la notion de « tri », qui relève pourtant de l'opération matérielle, ou encore à la « sélection » (Ducharme 2000, p. 19). Ce terme de *sélection* est utilisé dans plusieurs autres langues mais n'a pas toujours la même signification, ainsi en France elle désigne l'opération qui consiste à extraire des spécimens d'une série qu'on prévoit d'éliminer et s'oppose ainsi à l'échantillonnage (Favier, Neirinck 1993). Les problèmes terminologiques n'épargnent en rien la Suisse plurilingue et sans réelle tradition archivistique nationale. Ainsi les archivistes alémaniques utilisent tant *Bewertung* qu'*Überlieferungsbildung*, notion plus large qu'on peut traduire par « constitution du patrimoine archivistique », mais dont l'évaluation reste le cœur. Tandis que les tessinois usent la terminologie archivistique italienne de

[7] Que sont la création, l'évaluation, l'acquisition, la classification, la description, la diffusion et la préservation, définies par Carol Couture notamment dans *Les fonctions de l'archivistique contemporaine* (1999).

selezione e scarto – sélection et élimination (Burgy, Egli, Schmutz 2007, p. 282). Les termes d'évaluation, *selezione*, *appraisal* ou *Bewertung* portent tous en eux des connotations d'une certaine tradition archivistique mais visent en fin de compte le même but, soit décider que garder et que jeter.

1.3.3 Gestion électronique des affaires

En Suisse, la gestion électronique des affaires dans le domaine public est généralement associée au terme générique « GEVER », le concept de la Confédération pour la gestion électronique des affaires. GEVER se compose d'un système de classement structuré qui couvre toutes les tâches de l'unité administrative. Le système de classement implique la constitution des dossiers contenant tous les documents nécessaires à la traçabilité d'une affaire ainsi que les directives d'organisation, qui règlent la manière dont les collaborateurs gèrent les informations pertinentes pour un dossier et définissent le processus à suivre par les documents, de leur création à leur sortie du système et précise les types de documents qui ne sont pas pertinents et qu'il n'est par conséquent pas nécessaire d'enregistrer (Archives fédérales suisses 2022). En d'autres termes, GEVER met à disposition de la cyberadministration des fonctionnalités pour le *Records management* (Zeller 2013). Plus généralement, nous opposerons dans ce travail la gestion électronique des affaires de *type GEVER* aux environnements où aucune gestion unifiée n'a été mise en place.

1.3.4 Maturité de la gestion documentaire

La maturité de la gestion documentaire, intimement lié au concept de *Records management* et de la gestion électronique des affaires, est un aspect décisif dans l'approche de l'évaluation archivistique. En effet, les ressources (techniques et temporelles) nécessaires à une évaluation pertinente seront bien supérieures pour des documents issus d'un environnement d'une maturité faible, à l'inverse d'une documentation issue d'un environnement mature où une évaluation prospective a eu lieu et où des instruments de *Records management* sont en place et appliqués, qui demanderont beaucoup moins d'effort au moment de l'évaluation.

Le mandat réalisé par la HEG-GE distingue trois « niveaux de maturité ». Le niveau 1 correspond à un lot de données qui contient des fichiers en vracs sans informations descriptives ou très peu (pas de métadonnées). Le niveau de maturité 2 implique un plan d'archivage, mais qui n'est pas encore appliqué ou mal appliqué dans les dossiers réels. Le niveau de maturité 3 implique un lot de données qui contient un plan d'archivage et des dossiers d'activités bien formés (Makhlouf Shabou, Tièche 2018b, p. 3)

1.3.5 Collections de fichiers non structurées et vracs numériques

Dans le contexte numérique, le « vrac » se caractérise par un ou plusieurs des éléments suivants : peu ou pas contextualisé (producteur et/ou missions et projets non connus) ; peu ou pas organisé (notamment sans arborescence) ; peu ou pas identifié (nommage non signifiant *a priori*) ; avec des documents aux statuts différents et difficiles à établir (originaux et copies ; versions intermédiaires et versions finales/validées) ; avec des contenus, des formats et une volumétrie hétérogènes ; avec peu ou pas de métadonnées (de contexte, techniques, descriptives) ; souvent à prendre en charge dans des conditions particulières (sauvetage d'un fonds, départ d'un agent, transfert pour mutualisation, etc.) (Association des archivistes français 2018).

Dans la littérature germanophone on trouvera aussi le terme « *unstrukturierte Fileablage* »[8] ou encore « *Dateisammlungen* » (Lenartz 2020; Türck 2014; Schludi 2013; Puchta, Naumann 2017; Taylor 2016), traduit par le CECO (2022) comme « Collections de fichiers [non-structuré] ». Bien qu'il n'y ait pas encore de consensus autour de la définition exacte du terme (Puchta, Naumann 2017, p. 5), Stephan Lenartz propose la définition :

> *« Par collections de fichiers non structurées on entend une quantité de données numériques de type non déterminé, ne suivant aucun ordre prédéfini, déposé dans un système de répertoire et proposé aux archives dans leur intégralité, par exemple sur un support de donnée amovible portable. Si de telles collections sont également créées au sein des administrations en marge des systèmes de gestion des affaires, elles sont avant tout caractéristiques des archives privées. »*[9] *(2020, p. 16)*

1.3.6 Automatisation et intelligence artificielle

Dans la littérature consultée, les termes d'automatisation et d'intelligence artificielle (AI) apparaissent fréquemment mais sont peu définis. Concernant l'automatisation nous pouvons nous accorder sur la 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). Pour le domaine de la gestion documentaire, la *National Archives and Records Administration* (NARA) (2014, p. 10-14) distingue cinq approches (ou « degrés ») d'automatisation[10] :

- *No automation* – Pas d'automatisation, soit 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 processus et capturer les métadonnées nécessaires pour définir leur sort après leur durée d'utilité administrative et/ou légale ;
- *Modular Re-usable Management Tools*, des outils de gestion modulables et réutilisables : une approche intégrée qui fournit des outils, services ou applications de

[8] « *Ablage* » pouvant être traduit par « dépôt », « rangement » voire « rack » (LEO 2022)
[9] Traduction libre de l'auteur : « *Unter einer unstrukturierten Fileablage oder Dateisammlung wird eine Menge genuin digitaler Daten nicht näher bestimmten Typs verstanden, die in einer ebenfalls nicht näher definierten Ordnung in einem Verzeichnissystem abgelegt sind und dem Archiv in ihrer Gesamtheit, beispielsweise auf einem tragbaren Wechsel-Datenträger, angeboten werden. Solche Ablagen entstehen abseits der geregelten Aktenführung auch in Behörden, vor allem aber sind sie typisch für nicht-amtliche Überlieferungen, beispielsweise für digitale Nachlässe.* » (Lenartz 2020, p. 16)
[10] Traductions libres de l'auteur reprises du travail *Automatisation des fonctions archivistiques pour les données textuelles : quels outils et quelles fonctionnalités pour l'archiviste ?* Aurélie Bavaud, Sébastien Bischoff et Denis Bussard (2022).

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*, 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.

Basma Makhlouf Shabou précise : « L'automatisation ne doit pas être confondue, tout au moins, à deux processus : le processus de transposer les outils et les pratiques de gestion des archives papier aux archives électroniques (Bailey 2009, p. 93) et la systématisation qui consiste à concevoir et à formaliser le processus opérationnel d'une fonction. » (2015b, p. 200)

L'intelligence artificielle (AI) peut être définie de manière très générale comme toute technologie qui permet d'automatiser la connaissance humaine (McCorduck 2018). Cependant, sa nature en constante évolution en rend la définition difficile, à plus forte raison qu'une fois habitués à une technologie impliquant de l'AI, en général nous arrêterons de la considérer comme de l'AI. À l'instar d'autres technologies innovantes, l'AI dispose de son propre jargon (*expert systems, rule engines, machine learning, deep learning, neural networks,* etc.) et la confusion des statuts et de l'efficacité de ces technologies, parfois abandonnées puis retrouvées, rend l'appréhension particulièrement difficile au profane (Rolan et al. 2019, p. 181).

On distingue en général l'automatisation basée sur des règles (*rule-based*) et l'automatisation basée sur des modèles statistiques. De fait, il est en général fait référence à la deuxième catégorie quand le terme d'AI est utilisé et principalement pour désigner du *machine learning* (ML), soit le développement d'outils ou logiciels qui apprennent automatiquement sur la base de données. Si le ML a des applications multiples, dans notre travail, nous traiterons avant tout du *Natural Language Processing* – le traitement automatique du langage naturel (soit le langage humain), ainsi que la classification automatique des documents.

2. Méthodologie

Notre recherche appliquée est de type descriptive et qualitative dans la mesure où elle s'intéresse à décrire un processus et les concepts qui s'y réfèrent : l'évaluation aux sein des AEN, les dimensions d'évaluation et les fonctionnalités associées. L'évaluation de documents numériques étant toujours un phénomène en cours de maturation, notre recherche est exploratoire.

2.1 Collecte des données

La récolte de données s'est faite par différentes méthodes synthétisées dans le Tableau 1.

Tableau 1 : Instruments de collecte et données attendues

		Instruments de collecte				
		Entretiens	Recherche documentaire	Questionnaire	Documentation interne	Observations
Attentes	Prélèvement du *workflow* d'évaluation de documents analogiques	X			X	X
	Définition des besoins *ArchiSelect*	X			X	
	Éléments pour la proposition du *workflow* d'évaluation de documents numériques		X	X		X
	Possibilités fonctionnalités existantes		X	X		

2.1.1 Entretiens

2.1.1.1 Archives de l'État de Neuchâtel (AEN)

Plusieurs entretiens en groupe, de type semi-dirigé voire non-dirigé ont été menés à l'interne des AEN. Une feuille de route, ainsi qu'une grille d'entretien ont été préparés au préalable et la discussion durant l'entretien était menée par l'auteur. Ces entretiens ont été enregistrés et partiellement retranscrits, s'agissant de séances de travail internes, les retranscriptions ne sont pas mises à disposition. Deux entretiens d'une durée de respectivement trois et deux heures ont ainsi été menés en commun.

Le premier entretien, réunissant le « pôle évaluation » (voir 4.1) a eu lieu dans les locaux des AEN le 23 mai 2022 et avait comme objectifs principaux : de réaliser un cadre stratégique de l'évaluation ; de clarifier les pratiques d'évaluation ayant cours au sein des AEN ; et de clarifier certains aspects autour des fonctionnalités attendues par *ArchiSelect*. L'entretien s'est terminé sur une discussion autour du mandat réalisé par l'HEG-GE et spécifiquement autour des variables issues du cadre conceptuel (voir 3.2.1.1).

Le deuxième entretien a réuni les mêmes personnes, ainsi que le chef de projet *AENeas* et avait pour objectif de clarifier le périmètre précis d'*ArchiSelect*, notamment son interfaçage avec les autres outils de la *SuiteArchi*. Une discussion autour du questionnaire d'approbation des variables (voir 2.1.3) en a suivi.

Ces entretiens ont été complétés par plusieurs entretiens plus informels pour clarifications de certains points, notamment avec le gestionnaire d'information des AEN, Grégoire Oguey le 16 mai 2022 concernant le *workflow* d'évaluation de documents analogiques et concernant le mandat réalisé par la HEG-GE. Des séances de suivi avec l'archiviste cantonal et mandant du Mémoire, Lionel Bartolini ont eu lieu à intervalle régulier.

2.1.1.2 Archives cantonales de Suisse romande

Dans un premier temps, notre travail prévoyait de mettre un plus grand accent sur l'état des lieux de l'évaluation, non seulement à Neuchâtel, mais également en Suisse romande (excepté le canton de Berne). Des entretiens semi-directifs (grille d'entretien disponible en Annexe 1) ont ainsi été menés avec des personnes responsables de l'évaluation dans les archives cantonales romandes. Les objectifs du mémoire s'étant finalement recentrés autour des fonctionnalités soutenant l'archiviste dans sa pratique de l'évaluation de documents numériques, nous avons renoncé à une retranscription et une analyse spécifique de ces données. Si peu de renvois directs seront ainsi fait à ces entretiens, les données récoltées ont sous-tendus notre réflexion et ont permis de mieux positionner les spécificités du canton de Neuchâtel à l'échelle de la Suisse romande.

Nous résumerons tout de même ici quelques tendances générales recueillies :

- hormis le Valais qui dispose d'une *Politique d'acquisition* (Archives de l'État du Valais 2020), aucun canton romand ne dispose d'un document officiel définissant une politique d'évaluation ou des critères d'évaluation ;
- si la norme ISO 15489 sur le *Records Management* (2001) est évidemment connue par les archivistes, l'application pratique lors de l'évaluation archivistique de documents analogiques des qualités de *représentativité*, *valeur probante* et *exploitabilité* n'est pas systématique ;
- aucun canton n'a implémenté d'outil d'évaluation archivistique (rétrospective) pour les documents numériques, et aucun outil de soutien à l'évaluation n'a été cité ;
- l'avancée de l'évaluation prospective, soit la mise en place de calendriers de conservations est très disparate : Dans le canton du Jura, chaque service dispose d'un calendrier, tout comme dans le canton de Vaud où ils sont cependant jugés de qualité très diverses et un grand chantier de mise à jour et prévu, en Valais, Fribourg et Neuchâtel, la situation est variable.

2.1.1.3 Haute école de gestion de Genève

D'autre part, un entretien informel a été mené avec Arnaud Gaudinat, responsable de l'Axe 2 du mandat réalisé par la HEG-GE (voir 3.2.1.2), le 23 juin 2022 dans le but de clarifier quelques questions de compréhension techniques autour du mandat réalisé pour les AEN, notamment des aspects de *text mining*.

2.1.2 Recherche documentaire

Une recherche documentaire approfondie a été menée pour définir les principaux concepts entourant l'évaluation archivistique, notamment dans le cadre numérique ainsi que les technologies disponibles pour le soutien à l'évaluation. Notre recherche s'est ainsi basée sur les principales revues d'archivistique francophones, anglophones et en veillant à inclure également la littérature germanophone. Une attention particulière a été portée aux études de cas, ainsi qu'aux projets proposant des outils et *workflows* de traitement numérique des archives. De manière générale, nous avons privilégié des références récentes, les technologies évoluant rapidement dans les domaines techniques, notamment le *machine learning.* Divers travaux académiques issu de contextes locaux ont également été consultée

(Veuve 2021; Fritz 2021; Chevieux 2019; Tièche 2015; Mellifluo 2008). Toutes les références ont été gérées dans le logiciel *Zotero*.

Enfin, la documentation interne du projet *AENeas* (rapports, présentation de projet, cahier des charges d'*ArchiClass*) et les *livrables*[11] issus du mandat réalisé par la HEG-GE ont représenté une source d'importance majeure.

2.1.3 Questionnaire

Afin de mesurer l'approbation des variables (éléments mesurables sur les documents numériques) proposées par le mandat réalisé par la HEG-GE, un questionnaire en échelles de Likert[12] a été soumis aux archivistes du « Pôle évaluation » des AEN. Ce questionnaire, déjà soumis à un panel d'expert-e-s dans le cadre du mandat de recherche réalisé par la HEG-GE a été mis à disposition par l'équipe de la HEG-GE, avec le grand avantage qu'il avait déjà été pré-testé et que le protocole de réalisation pouvait être repris tel quel. Les résultats de ce questionnaire sont discutés dans la partie 5.3.1.2, les résultats bruts sont disponibles en Annexe 2.

2.1.4 Observations

L'auteur de ce travail étant également archiviste aux AEN et pratiquant l'évaluation lors de la réalisation de ce Mémoire, des observations, bien que formalisées uniquement en prise de notes, ont été une source d'information de premier plan. En effet, la période de récolte des données coïncidant avec la période de formation à l'évaluation aux AEN, un regard externe et neutre a pu être posé sur les pratiques d'évaluation. La période de formation soulevant questions et discussions entre collègues, celles-ci ont pu être mis à contribution et intégrées notamment dans la modélisation du *workflow* d'évaluation de documents analogiques (voir 5.1.1). Ce double-rôle peut bien entendu induire certains biais, contre lequel nous nous sommes prémuni, du mieux possible, par une posture réflexive et une introspection critique continue (Fortin 2016, p. 174).

2.2 Périmètre de l'étude

Dans ce travail nous traitons des fonctionnalités pouvant soutenir l'archiviste dans la fonction d'évaluation archivistique dans le domaine numérique tel que prévu dans le périmètre d'*ArchiSelect.* Nous laisserons de côté en grande partie l'évaluation prospective ainsi que la description des documents dans un but de diffusion.

Nous laisserons également de côté, pour une question de faisabilité, toute réflexion autour de l'interface graphique, travail entamé par le mandat réalisé par la HEG-GE (voir 3.2.1.3). Une réflexion approfondie autour des métadonnées a également été écartée du périmètre de ce Mémoire, réflexion largement entamé dans le cadre du mandat de la HEG-GE (Axe 2). Nous avons aussi décidé de sortir du périmètre de ce Mémoire toutes les fonctionnalités « administratives » du logiciel, celle-ci étant notamment dépendantes des choix d'implémentation des fonctionnalités de soutien à l'évaluation.

[11] Les résultats du mandat de recherche ont été restitués sous forme de *livrables*, soit des documents ou rapports répondant aux objectifs fixés par le mandat, les résultats et la structure de ces *livrables* sont définis dans la partie 3.2.1.

[12] « Échelle d'attitude constituée d'une série d'énoncés déclaratifs pour lesquels le répondant exprime son degré d'accord ou de désaccord » (Fortin 2016, p. 333)

Enfin, nous précisions que toutes les questions liées spécifiquement à l'évaluation de document non textuels ou hybrides, tel photos, vidéos, bases de données ainsi que l'archivage d'internet et autres réseaux sociaux ont également été sortis du périmètre pour des questions de faisabilité.

3. État des lieux de l'évaluation archivistique à l'ère du numérique

Dans cette partie, nous réalisons un état des lieux de l'évaluation et plus particulièrement l'évaluation de documents numériques. Pour ce faire, nous réalisons une revue de littérature approfondie puis présenterons les principaux projets autour de l'archivage numérique en mettant l'emphase sur la fonction de l'évaluation.

3.1 Revue de littérature

3.1.1 Évaluation archivistique

3.1.1.1 Courants théoriques majeurs

Des panoramas complets des théories de l'évaluation ont été réalisés par divers auteur-e-s, notamment par Carol Couture (1996 ; Ducharme, Couture 1996) Basma Makhlouf Shabou (2010 ; 2015b) et plus récemment par le prisme de la dispute entre positivistes et postmodernistes au sein de l'archivistique par William Yoakim dans le cadre de sa thèse (2022).

Pour replacer notre réflexion dans un cadre historique plus large, nous évoquerons brièvement les auteur-e-s et concepts qui ont façonné l'évaluation archivistique. Si le développement des pratiques de l'évaluation a essentiellement eu lieu durant les années 1970 (Makhlouf Shabou 2015b), c'est dès la fin du 19e siècle que des manuels sont rédigés à l'attention des archivistes : par Natalis de Wailly appelant au respect de provenance en 1841, puis par les archivistes hollandais (Muller, Feith, Fruin 1910) dont les principes sont repris en grande partie par l'archiviste britannique Hilary Jenkinson (1937) qui, suite à la première guerre mondiale et un accroissement de la masse documentaire, veut remettre la responsabilité de l'évaluation sur le producteur qui en serait le meilleur juge reléguant ainsi l'archiviste au rôle de gardien neutre ne pratiquant ni d'élimination, ni de changements dans les fonds.

Après la deuxième guerre mondiale et en période de *New Deal*, les États-Unis font face à une explosion de la production documentaire d'une autre ampleur, la commission Hoover souhaite ainsi réformer les instances gouvernementales et prône une élimination progressive des documents. On mise dans la tradition étasunienne sur l'élaboration de critères : rationnels – importance donnée au document par le créateur-ice, importance pour l'histoire administrative et valeur historique (Brooks 1940) ; ainsi que l'implication du coût engendré par les archives définitives (Bauer 1946). Enfin par la théorie des valeurs formulée par Theodore Schellenberg (1956 ; 1965) qui va s'imposer en maître à penser.

Dès les années 1970 on plaide en Allemagne pour une approche sociétale de l'évaluation (B ooms 1972; 2001), les archives doivent être le reflet de l'activité de la société entière et l'opinion publique doit prédominer. Ainsi, l'organisme créateur et l'utilisation des archives deviennent des facteurs importants dans le cadre de l'évaluation. La publication de la traduction du texte de Booms en 1987 a eu un impact important en Amérique du Nord (Couture 1996, p. 8), les principes sont repris par Helen W. Samue ls (1986a; 1986b) et Terry Eastwood qui développe l'idée de l'approche centrée utilisateur (1992). Frank Boles et Julia Marks Young (1991) réalisent un important travail sur les critères. Le résultat est certes difficile d'application mais est construit à partir de méthodes quantitatives « rassurantes » et exerce une grande influence sur les archivistes nord-américains (Ducharme 2000, p. 18).

Terry Cook (1992) apporte une avancée majeure en théorisant le concept de « macro-évaluation », une approche *top-down* de l'évaluation. À cette approche, Carol Couture, se réclamant d'une « archivistique intégrée » (1999, p. 108), ajoute le concept de « micro-évaluation » qui serait son « contraire apparent » (1996, p. 4), les deux formant en réalité un tout.

L'évaluation se base selon Couture sur la théorie des valeurs de Schellenberg (1956; 1965) ainsi que sur cinq principes directeurs : les archives témoignent de l'ensemble des activités de la société ; l'archiviste fait preuve d'objectivité et de contemporanéité dans le jugement qu'il porte ; l'archiviste respecte les liens qui unissent l'évaluation et les autres interventions archivistiques ; l'archiviste respecte l'équilibre entre la finalité administrative et patrimoniale de son intervention ; ainsi que l'équilibre des considération du contexte de création des archives et celles liées à leur utilisation (Couture 1996, p. 17-18).

Enfin, Basma Makhlouf Shabou mène des recherches autour de la qualité des archives définitives, de leur mesure et du manque de normalisation (2010; 2012a; 2015a; Makhlouf Shabou et al. 2020) qui seront discutés dans la partie 5.3.1 de ce Mémoire.

3.1.1.2 L'évaluation archivistique en Suisse

Le paysage archivistique suisse se caractérise par une « diversité surprenante pour un territoire si exigu » (Burgy, Roth-Lochner 2002). L'ouvrage collectif *Pratiques archivistiques en Suisse* (Coutaz et al. 2007) donne un excellent aperçu historique du paysage archivistique suisse et de ses pratiques, un chapitre y est d'ailleurs voué à l'évaluation (Burgy, Egli, Schmutz 2007).

La réflexion de la communauté professionnelle archivistique helvétique se limite dans un premier temps à rédiger, dès les années 1950, des comptes rendus. C'est dans les années 1980, sous la pression des masses documentaires à traiter, qu'apparaissent les premières expressions concrètes de l'évaluation (Coutaz 2011). L'évaluation est alors essentiellement fondée sur l'expérience professionnelle, et s'appuie notamment sur l'expérience d'historien-ne des archivistes. Les critères appliqués sont d'ailleurs largement basés sur les besoins futurs de la recherche historique (Burgy, Egli, Schmutz 2007, p. 285).

Une première journée d'étude de l'Association des Archivistes suisses (AAS) consacrée à l'évaluation en 1995 suscitera une série d'articles référence germanophones dans la revue *Arbido* (Haener 1995; Zweifel 1995; Zwicker 1995; Bütikofer 1995; Roth-Lochner 1995). Par la suite une première thèse est publiée (Halbeisen 1999). Dix ans après la journée d'étude de l'AAS, Josef Zwicker signe un autre texte de référence en proposant un état des lieux de la discussion sur l'évaluation en Suisse (2005).

Pour éviter de donner libre cours à la « fureur du particularisme » (Burgy, Roth-Lochner 2002), l'AAS a mis en place le groupe de travail « Évaluation »[13] – ultérieurement appelée la Commission de coordination (*KoKo*) – qui fonctionne comme une structure d'échanges, d'élaboration de modèles et de veille. Le groupe offre des recommandations d'archivage à l'intention des Archives fédérales et cantonales, et structure et centralise les débats autour de

[13] https://vsa-aas.ch/fr/association/groupes-de-travail/evaluation/ [Consulté le 28 juillet 2022]

l'évaluation (Coutaz 2011). La plateforme du groupe de travail recense par ailleurs les travaux académiques suisses en lien avec l'évaluation[14].

Si la thématique du numérique avait été agendée à la journée de travail de l'AAS de 1995 sur l'évaluation, encore considérée comme une problématique « futuriste » (Zeller 2003), il existe à notre connaissance peu d'articles suisses ayant fait référence dans le domaine. Jean-Daniel Zeller (2013) propose tout de même un historique du développement de l'archivage électronique en Suisse (décennie 2002-2012).

Enfin, dans un contexte plus régional et en lien direct avec notre travail, Nils Veuve (2021) a réalisé un Mémoire autour d'*AENeas*, le concept d'archivage numérique du canton de Neuchâtel, tandis que Grégoire Oguey et Pascal Schneiter (2018) ont publié dans *Arbido* un article sur le logiciel *ArchiSelect* spécifiquement.

3.1.2 L'évaluation archivistique de documents numériques

3.1.2.1 Les débuts

Bien que l'évaluation de documents numériques puisse sembler un phénomène récent, un groupe de travail de la *National Archives and Records Administration* se serait formé dès 1978 « pour examiner si les données lisibles par machine avaient une valeur historique » (Thibodeau 1998). Les premières recommandations, émises la même année, proposent déjà que l'évaluation de documents numériques se fasse sur la base d'un calendrier de conservation (Dollar, 1978), tout comme la problématique de l'obsolescence des formats, de l'importance des caractéristiques techniques des supports informatiques et les coûts de préservation (Naughler 1983).

Durant les années 1990 apparaît le *post-custodialism* qui défend l'idée que les entités créatrices aient la « garde » (*custody*) des documents numériques qu'ils ont produits, de peur que les archivistes n'aient pas les ressources nécessaires à la gestion complexe des documents numériques (Ham 1981; Bearman 1991; O'Shae, Allen 1996). Bearman devra faire face à certaines critiques, notamment du fait que l'idée de *post-custodialism* est très peu compatible avec des archives d'origine privées, l'idée a toutefois fourni des arguments à Terry Cook pour l'élaboration du concept de macro-évaluation (1994).

Au milieu des années 1990 Le projet *INTERPares* réuni autour de Luciana Duranti, se positionne contre la théorie post-custodialiste (Marsden 1997; Cunningham 1996), plaidant justement pour une forte « garde » des documents numériques par les archivistes qui doivent à tout prix préserver l'authenticité des documents inactifs (Duranti, Macneil 1996).

Pour une revue complète des débuts de l'archivistique à l'ère du numérique, voir David Rajotte (2010).

[14] Travaux de la Haute école de gestion de Genève (HEG-GE), section Science de l'information ; de l'Université de Berne et Lausanne, *Weiterbildungsprogramm Archiv-, Bibliotheks- und Informationswissenschaft; Master of Advanced Studies in Archival, Library and Information Science (MAS ALIS)* ; et la *Fachhochschule Graubünden, Hochschule für Technik und Wirtschaft HTW Chur, Informationswissenschaft.* https://archiv.vsa-aas.ch/wp-content/uploads/2021/12/Diplomarbeiten_Bewertung_2021_11.pdf [Consulté le 28 juillet 2022]

3.1.2.2 Tendances actuelles

Si nous avons vu que l'évaluation archivistique de documents numériques n'est pas un phénomène récent, il est d'autant plus étonnant de voir le peu d'assise théorique et de solution techniques normalisées existantes (Lenartz 2020, p. 17). Un certain consensus règne sur le fait que l'évaluation de documents numériques ne diffère pas, sur le principe, de l'évaluation de documents analogiques (Huber 2009; Sloyan 2016; Béchard, Fuentes Hashimoto, Vasseur 2020), Cependant, deux considérations rendent l'évaluation de matériaux numériques substantiellement différente de l'analogique : « *digital materials exist at multiple levels of representation and […] they are directly machine readable* » (Lee 2018).

La théorie archivistique de l'évaluation reste donc valide et les critères les mêmes, les *processus* d'applications diffèrent cependant (Sloyan 2016), notamment du fait que les documents numériques offrent d'autres métriques et angles d'approche que les documents analogiques (Wettman 2008). Le véritable défi est ainsi de traduire ces principes et critères dans le numérique tout en différenciant bien leur *archivabilité* et *valeur archivistique* (Türck 2014).

Les textes théoriques et récents sur l'évaluation de documents numériques faisant référence, sont, à notre connaissance, rares voire inexistantes (pour un état des lieux de la question, voir Frank Bischoff (2014)). Ainsi, la bibliographie recèle avant tout des études de cas. Si nous ne pouvons que saluer une approche pratique de la problématique, cela est également représentatif d'un état de la recherche en cours de maturation.

Ainsi, diverses études relatent des retours d'expérience de *traitement* de collections de fichiers peu structurés et les difficultés rencontrés (Mumma, Dingwall, Bigelow 2011). Il est d'ailleurs intéressant de relever que dans le cadre de ces expériences, largement menées en milieu anglophone, on préfère les termes d'*analysis*, *review* ou *identification* et de *digital processing,* évitant l'utilisation du terme d'*appraisal* (évaluation) (Belovari 2017, p. 56) voire laissant de côté l'étape de l'évaluation, se contentant d'une déduplication (Shein 2014).

Belovari (2017) mène ainsi une étude de cas avec une approche *More product, Less Process* (MPLP)[15] au sein des archives d'État de Ludwigsburg (Allemagne). Partant d'une collection de fichiers de 677 GB, Belovari crée un *workflow*, en utilisant des logiciels commerciaux et accompagne sa démarche par une réflexion méthodologique (concernant le traitement du fonds particulier de cette étude de cas, également voir Naumann (2017) et Knobloch (2019)).

3.1.2.3 Automatisation, *text-mining* et *machine learning* appliqué à l'évaluation archivistique

3.1.2.3.1 Automatisation

Si on évoque l'automatisation des fonctions archivistique depuis les années 1960 (Bunn 2016), son opérationnalisation, du moins partielle, a pris une autre tournure avec l'apparition – et surtout la démocratisation – de l'intelligence artificielle (AI), soit essentiellement l'utilisation du *text mining*[16] et du *machine learning*. Certains auteur-e-s appellent ainsi à l'utilisation des

[15] À ce propos, également voir Mark A. Greene et Dennis Meissner (2005) et Greene (2010)

[16] Outre une automatisation de certaines fonctions documentaires, le *text-mining* entraîne également un changement de paradigme dans l'utilisation des archives par l'application du concept de *distant reading* (Moretti 2013) qui voit l'entier de la production littéraire comme un seul corpus : un fonds ne sera plus vu uniquement comme une collection de textes, mais comme des données auxquelles il faut *donner du sens* (Moss, Thomas, Gollins 2018, p. 120).

nouvelles technologies des géants du web pour la gestion documentaire (Bailey 2009). Nous passerons ici en revue certains textes traitant de ces technologies appliquées dans les projets archivistiques, sans avoir la prétention de fournir une revue de littérature approfondie sur le *machine learning*[17].

De premières expériences sont ainsi menées dans les années 2000 (Kim et al.2006) et la *National Archives and Records Administration* (NARA) publie un rapport (2014) qui marque un jalon (Hooland, Coeckelbergs 2018), soulignant le besoin d'automatisation et proposant quelques pistes d'application, mais restant toutefois dans un certain niveau d'abstraction. Par la suite, divers articles enquêtent sur les exigences que demanderait une telle automatisation (Elragal, Päivärinta 2017) et spécifiquement les défis soulevés par l'évaluation (Harvey, Thompson 2010; Makhlouf Shabou 2015b).

Plusieurs articles récents fournissent de bons tours d'horizon et états des lieux de l'utilisation de l'AI dans l'archivistique et applications futures possibles (Rolan et al. 2019; Fiorucci et al. 2020; Colavizza et al. 2022). Tous concluent grossièrement par le même constat : beaucoup de progrès ont été réalisés, cependant, l'utilisation concrète de l'AI sur les archives reste rare et la transformation des expérimentations en une infrastructure durable va demander énormément de moyens. L'implication qu'auront ces changements pour le métier d'archiviste est régulièrement abordé, Marciano et al. (2018) plaident ainsi pour la création d'un nouveau corps de métier interdisciplinaire entre archivistique et informatique : la *Computational archival science*.

Diverses études de cas sont ici également à signaler, à commencer bien entendu par la recherche menée par Makhlouf Shabou et al. (2020) dans le cadre du mandat de recherche réalisé pour les AEN traité dans la partie 3.2.1 de ce travail. Stephan Lenartz (2020), propose une étude de cas automatisant des séquences entières du *workflow* d'évaluation d'une collection de fichiers contenant notamment des photos à l'aide d'une série de scripts *python*. Une proportion élevée d'études de cas se penche sur l'évaluation de courriels (Vellino, Alberts 2016; Alberts, Forest 2012; Alberts, Vellino 2013; Vinh-Doyle 2017; Schneider et al. 2019; Gilliland 2016; Decker et al. 2021; Drake 2015).

3.1.2.3.2 Natural Language Processing (NLP)

Une utilisation spécifique du *machine learning* éveille particulièrement l'attention dans le domaine de l'archivistique, notamment pour son utilisation dans l'évaluation : le *Natural Language Processing* (NLP) (aussi voir 5.3.2.4). L'utilisation du NLP en archivistique est issue du domaine des *Digital Forensics* (criminalistique informatique) qui connaît certaines similarités avec l'évaluation de documents numériques : comme lors d'une enquête criminelle, l'archiviste doit investiguer le contenu de matériaux qui lui a été proposé, souvent incomplets et partiels (Goodman 2019). Le rapprochement entre les métiers patrimoniaux et les *digital forensics* est thématisé notamment par Kirschenbaum et al. (2010) et l'idée d'utiliser des outils d'investigation est évoquée déjà ultérieurement, notamment par Samuel Ross et Ann Gow (1999). Les outils de *digital forensics* ne sont cependant pas forcément adaptés aux besoin

Ce changement de paradigme aurait également de grandes implications sur le métier d'archiviste et sur l'évaluation.

[17] D'après Arnaud Gaudinat (entretien du 23 juin 2022 réalisé à distance), un tel état des lieux devrait être réalisé avant l'implémentation de fonctionnalités ayant recours au *machine learning* par une personne ou institution experte dans le domaine.

des archivistes (Kirschenbaum et al. 2010), le projet *BitCurator*, que nous présenterons dans la partie 3.2.2.3 cherche justement à rendre applicables les outils de *digital forensics* à l'archivistique (Lee, Woods 2017; Lee 2018; Goodman 2019).

Young et al. (2017) proposent une revue des tendances du NLP basée sur le *deep learning* spécifiquement, tandis que Tim Hutchinson (2020) fournit une vue d'ensemble des projets dans le domaine de l'évaluation archivistique. Il observe que la plupart des implémentations de NLP et de *Machine learning* sont basées sur des projets et encore au stade expérimental, il conclut en proposant cinq principes que devraient suivre le les outils d'évaluation utilisant le NLP : *usable, interoperable, flexible, iterative, configurable* (2020, p. 166-168). Morgan Goodman (2019) examine plus en détail l'outil de *Topic modeling* créé par *BitCurator NLP*. Le *Topic modeling* est également utilisé sur de volumineux corpus historiques, ainsi Chaney et al. (2016) mènent une étude sur la classification d'époques historiques déterminées à l'aide du langage présent dans les documents officiels. D'autres études explorent l'utilisation spécifique de la reconnaissance d'entités nommées (*Named entity recognigion* – NER) pour l'exploitation d'archives définitives (Hengchen et al. 2016; Hooland, Coeckelbergs 2018).

Enfin, différents auteur-e-s plaident pour le développement de la *sensitivity review*, soit l'identification automatique de données sensibles pour rendre accessible au public certaines archives numériques (Sloyan 2016; Moss, Gollins 2017; Jaillant, Caputo 2022) notamment à l'aide du NLP (Gollins et al. 2014). Schneider et al. (2019) testent ainsi l'outil *ePADD* pour l'évaluation de courriels incluant une fonctionnalité de *sensitivity review* basée sur le NLP.

3.2 Principaux projets d'archivage numérique

Cette partie du travail présente les principaux projets d'archivage numérique, en se concentrant sur les projets autour de l'évaluation et ses essais d'automatisation. Nous commencerons par présenter le mandat de recherche mené par la Haute école de gestion de Genève (HEG-GE) pour les Archives de l'État de Neuchâtel (AEN) déjà évoqué, avant de survoler les principaux projets internationaux et en finissant sur les quelques propositions suisses.

3.2.1 *Modèle d'évaluation systématique des documents : Concept & Preuve de concept*

Le mandat de recherche attribué par les AEN à la Haute École de Gestion de Genève (HEG-GE) mené par Prof. Basma Makhlouf Shabou et Prof. Arnaud Gaudinat sera discuté plus en détail pour la raison évidente qu'il traite de l'outil *ArchiSelect*.

Ce travail, organisé en trois axes, propose une réflexion autour des métriques archivistiques pour l'évaluation dans le premier[18] ; une preuve de concept de l'utilisation de la fouille de données dans le deuxième[19] ; et dans le troisième une traduction en interface graphique.

Les principaux résultats théoriques de cette recherche ont également été repris dans l'article *Algorithmic methods to explore the automation of the appraisal of structured and unstructured digital data* (Makhlouf Shabou et al. 2020).

3.2.1.1 Axe 1 : Dimensions d'évaluation

L'Axe 1 propose un cadre conceptuel structuré autour de « dimensions d'évaluation » (DEVs) basées sur la littérature académique et professionnelle, notamment la norme ISO 15489 sur le *Records management*. Cette norme précise que les qualités ou caractéristiques que doivent posséder les documents d'activités sont *l'authenticité*, la *fiabilité* et *l'exploitabilité* afin de constituer une preuve des événements de l'activité ou des opérations (ISO 15489 :2016, p. 4).

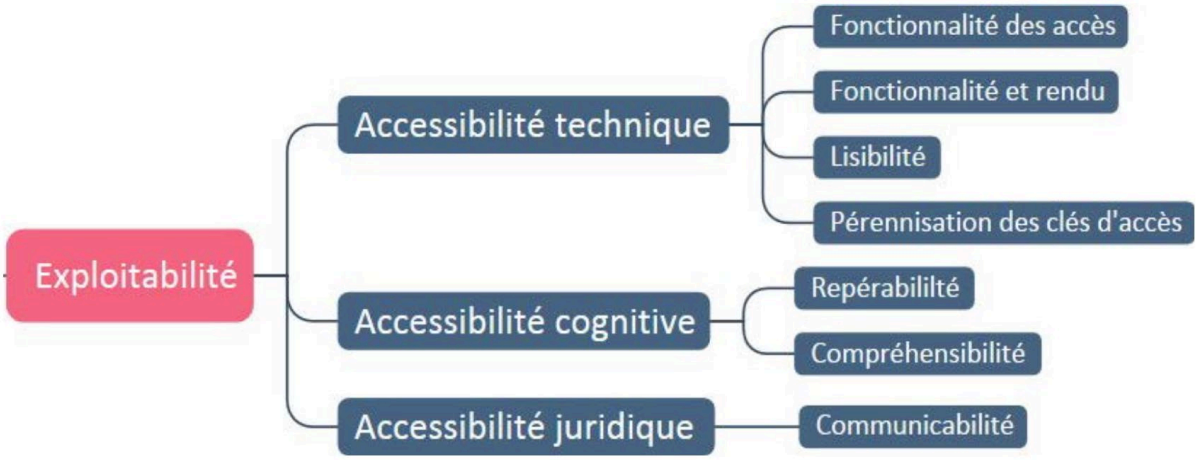
De ces trois dimensions ont ensuite été tiré huit « sous-dimensions » (DEV2), développées en vingt sous-dimensions (DEV3) (voir Figure 2 pour la dimension d'*exploitabilité*). Pour les opérationnaliser (soit les mesurer sur les documents numériques à évaluer), les auteur-e-s du rapport en ont tiré des variables – « véritables éléments de mesures, qui permettent de quantifier leur valeur associée » (Makhlouf Shabou, Tièche 2018a, p. 8). Une liste complète des DEVs et des variables, ainsi que leur définition sont disponibles dans l'Annexe 3. Dans la documentation il est mis en avant l'aspect modulable, adaptable et qui tend vers l'automatisation de ces variables. Parmi d'autres critères[20], celui de la capacité à l'automatisation est ainsi indiqué pour chaque variable, la mesure des variables peut ainsi être soit : *automatisable* (totalement applicable par la machine ; *semi-automatisable* (partiellement applicable par la machine) ; *manuelle et systématique* (totalement applicable de façon formelle par l'humain) ; ou *manuelle et subjective* (totalement applicable de façon informelle par l'humain) (Makhlouf Shabou, Tièche 2018b, p. 3).

[18] Dans un souci de lisibilité, il sera fait référence au « mandat HEG-GE » pour désigner l'ensemble de la documentation et les documents seront désignés en précisant l'axe et le numéro de livrable, « Axe 1, L3 » pour le *Livrable 3* de l'axe 1 par exemple.
L'Axe 1 est composé de trois livrables, répartis en quatre documents : *Livrables 1.1 & 1.2 : Cadre conceptuel et typologie des critères* (Makhlouf Shabou, Tièche 2018a) ; *Livrables 2.1 & 2.2 : Modèle opérationnel détaillé des métriques archivistiques et Nomenclature des critères et métriques d'évaluation* (Makhlouf Shabou, Tièche 2018b) ; *Livrable 3.1 : Études de cas* (Makhlouf Shabou, Tièche 2018c) ; *Livrable 3.2 : Concept et preuve de concept de métriques pour l'évaluation archivistique: Définition et validation de sets de métriques : Rapport de recherche* (Makhlouf Shabou, Tièche 2018d).

[19] L'Axe 2 est composé de six livrables, répartis en huit documents : *Livrable 1 : État de l'art* (Gaudinat, Knafou 2017a) ; *Livrable 2 : Description des données et propositions d'indicateurs clés possibles* (Gaudinat, Knafou 2017b) ; *Livrable 3 : Analyse des métadonnées externes aux dossiers d'activité* (Gaudinat, Knafou 2018b) ; *Livrable 4 : Preuve de concept de fouille de données* (Gaudinat, Knafou 2018c) ; *Livrable 5 : Test de faisabilité du modèle opérationnel* (Gaudinat, Knafou 2018d) ; *Livrable 5-3 : Résultat de l'enquête auprès des archivistes* (Gaudinat, Knafou 2018e) ; *Livrable 6 : Spécification, fouille de donnée - Rapport final* (Gaudinat, Knafou 2018a) ; *Annexes Livrable 6* (Gaudinat, Knafou 2018f).

[20] Critères d'exclusivité ; critères intrinsèque et extrinsèque ; critères relatifs au niveau de maturité de gestion documentaire ou d'applicabilité, voir (Makhlouf Shabou, Tièche 2018b, p. 3)

Figure 2 : Dimension d'évaluation d'Exploitabilité



(Makhlouf Shabou, Tièche 2018a, p. 8)

Les auteur-e-s du mandat ont soumis les variables à un panel d'expert-e-s en archivistique sous forme d'un questionnaire en échelles de Likert afin de valider leur pertinence. Pour tester la faisabilité du modèle opérationnel, trois études de cas ont été menés pour explorer les différents scenarios possibles de l'application des métriques.

3.2.1.2 Axe 2 : *Data mining*

L'Axe 2 du mandat de recherche présente l'approche « fouilles de données » (*data-* ou *text mining*) dans l'optique d'un traitement automatique des dossiers d'activité et l'intégration des métriques d'évaluation identifiés dans l'Axe 1. L'objectif principal étant la réalisation de spécifications pour l'outil *ArchiSelect*.

3.2.1.2.1 Métadonnées

L'Axe 2 réalise un *mapping* des métadonnées « internes » des types de fichiers susceptibles d'être trouvées dans un vrac à archiver (*Axe 2, L2*), *mapping* réalisé sur deux jeux de données représentatifs à l'aide de la suite *OpenSearchServer* (OSS)[21]. Un travail est également mené sur les métadonnées « externes » aux données (dossiers d'activité ou lot de données), en se focalisant sur les données provenant d'*ArchiClass* et des dossiers d'activité (*Axe 2*, *L3*).

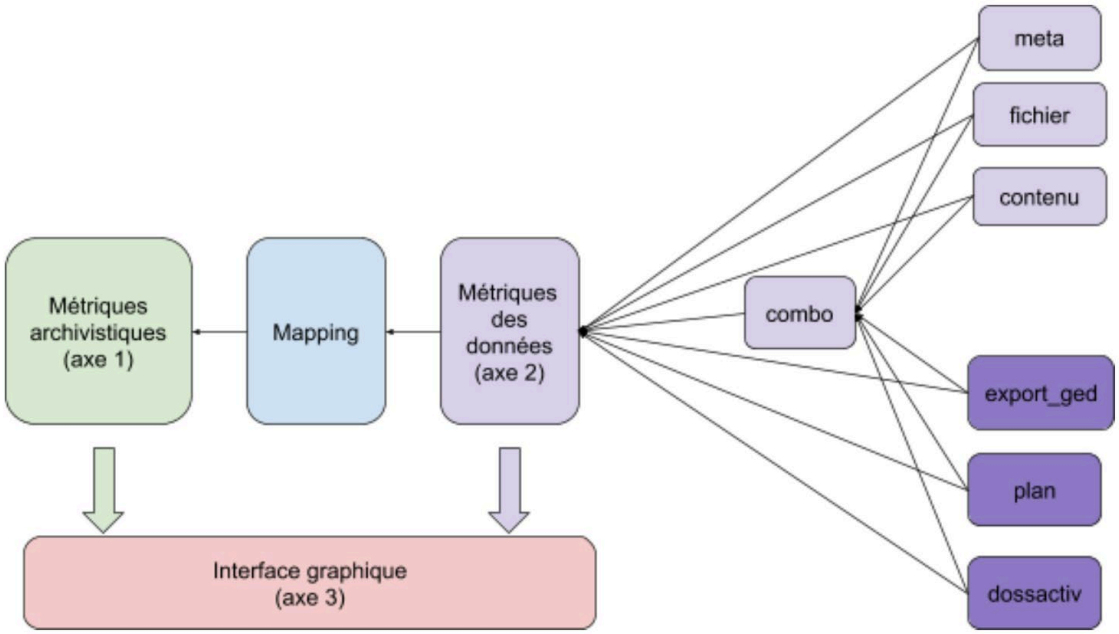
3.2.1.2.2 Mapping des métriques archivistiques et métriques de fouille de donnée

L'équipe de la HEG-GE propose un *mapping,* ou une ébauche de correspondance, entre les métriques archivistiques (les DEVs et variables développées dans l'axe 1, en vert sur la Figure 3) et les métriques de fouille de données[22] (en violet) (*Axe 2, L3*).

[21] Il s'agit d'un moteur de recherche *open source* offrant un jeu de fonction de recherche, d'indexation et *crawling*. https://www.opensearchserver.com/ [Consulté le 16 juillet 2022]. Les auteurs du Livrable précisent qu'OSS est basé sur l'API *Lucene*.

[22] Le concept distingue différents types de métriques de fouille : Le type « *meta* » qui correspondent aux métadonnées renseignées par les applications fichiers (par exemple *Word* ou *Excel* ; de type « *fichier* », renseignées par le système de fichiers ; de type « *contenu* » qui découlent en général d'analyse par des techniques de fouilles de données (et *Natural Language Processing*) ; les métriques de type « *politique* », provenant d'une analyse automatique d'une politique de gestion des documents (si disponible) écrite en texte libre (langage naturel). ; les métriques de type « *plan* » qui se réfèrent aux données extraites du plan d'archivage provenant du logiciel *ArchiClass* ; de type « *dossactiv* », qui se réfèrent aux

Figure 3 : Modèle de métriques et correspondance (*mapping*) entre les axes 1 et 2



(Gaudinat, Knafou 2018b, p. 3)

Les auteur-e-s du mandat tentent ainsi de mettre en relation les métriques archivistiques et les métriques de données, une correspondance exprimée sous forme de *règle* (Figure 4) qui combinent conditions et boucles sur les métriques, dont le résultat indique une note.

Dans le *Livrable 3* de l'Axe 2 sont ainsi proposées des ébauches de correspondances, soit des *règles* pour vingt variables, sélectionnées sur des critères d'automatisation potentielle qui sont à une exception près, issues des dimensions d'évaluation de *valeur probante* et *exploitabilité*. Nous donnerons ici l'exemple pour la variable 29 « Description du producteur » (Figure 4).

métadonnées potentiellement importées par des dossiers d'activité ; et de type « *combo* » qui ont pour but d'unifier les provenances de la même métadonnées pour associer des degrés de confiance à des valeur possible. (Gaudinat, Knafou 2018b, p. 4-5).

Figure 4 : Tableau des correspondances de règles pour la variable 29 « Description du producteur » des métriques archivistiques

règle (pseudocode)	note

`auteurs = list(meta_auteur, fichier_auteur)`
`for auteur in auteurs:`
`  if auteur in contenu_rec_nom:`
`  if auteur in nom_referenciel:`
	1

`auteurs = list(meta_auteur, fichier_auteur)`
`for auteur in auteurs:`
`  if auteur in contenu_rec_nom:`
	1/2

`auteurs = list(meta_auteur, fichier_auteur)`
`for auteur in auteurs:`
`  if auteur not in contenu_rec_nom:`
	0

(Gaudinat, Knafou 2018b, p. 19)

Dans l'exemple de la variable 29 « Description du producteur », le *pseudocode* définit en se basant sur des métadonnées choisies si une description du producteur est disponible et accorde une note de 1, ½ ou 0.

3.2.1.2.3 Preuve de Concept fouille de données et spécifications fonctionnelles

L'Axe 2 fournit une Preuve de Concept (*L4*) de l'utilisation de la fouille de données se basant sur un corpus de document mis à disposition par les AEN, ainsi qu'une collection de courriels interne à l'HEG-GE. Y sont analysés (extensions, nombre de métadonnées par document, distribution des extensions, production de documents au travers du temps) et indexés[23] les documents du jeu de données proposé par les AEN et les métadonnées utiles au calcul des métriques de données extraites. Trois métriques archivistiques ont ensuite été mesurées sur le jeu de données, sur la base des ébauches des *règles* présentés dans la section précédente. Ainsi, à titre d'exemple, pour la métrique « complétude des métadonnées », 64.9% des fichiers inspectés ont reçu la note de 1 soit la note maximale, 33.2 % ont reçu la note 0.5 et 1.9% la note 0 (Gaudinat, Knafou 2018c, p. 14).

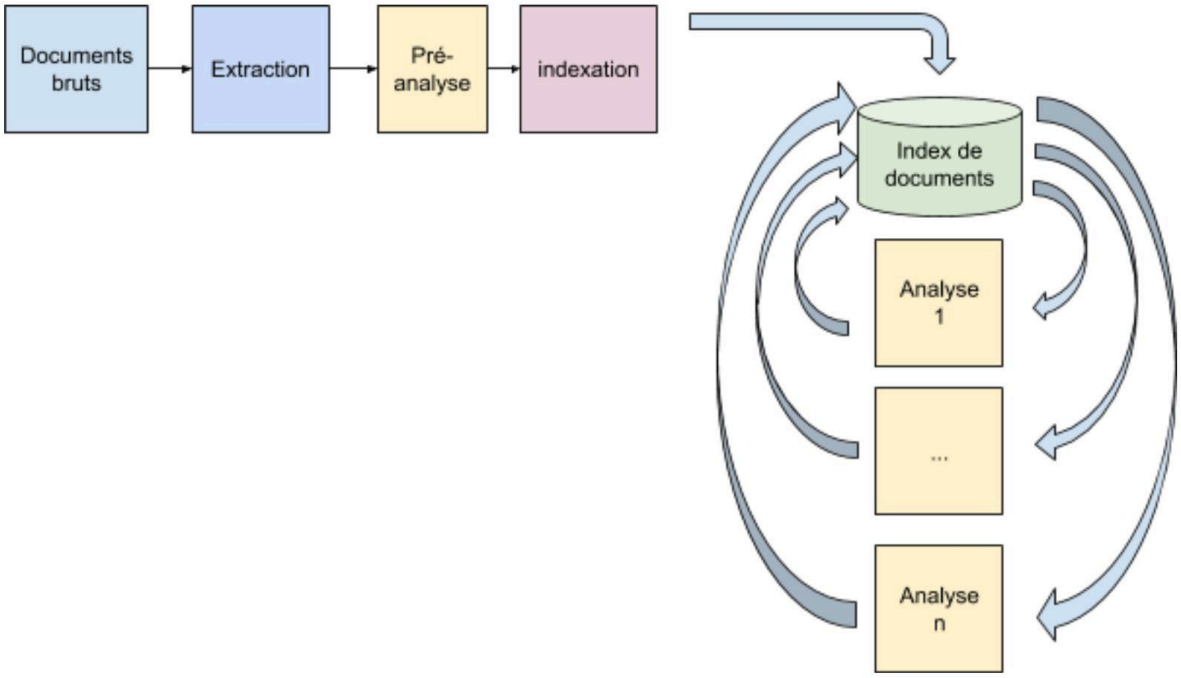
Une preuve de concept pour la reconnaissance d'entité nommée – ou *Named Entity Recognition* (NER) a également été fournie. La NER servira dans les métriques dites *combo* utilisées pour accorder plus de confiance aux métadonnées d'auteur d'un fichier. Enfin, une preuve de concept pour la classification automatique de documents a été réalisée avec l'exemple du document type « procès-verbal » en utilisant un algorithme de classification supervisée. Enfin, une preuve de concept pour la recherche de document similaire a été réalisée. Cette fonctionnalité permettrait d'identifier des documents similaires ou quasi-similaires par un rapport à un document donné, ce qui pourrait être utilisé dans l'indexation de documents similaires et pour une recherche dynamique de documents.

[23] Pour des questions de système d'exploitation ce n'est cette fois pas *OpenSearchServer*, produit *open source* plus apte à fonctionner sur des environnement type *Unix* (Gaudinat, Knafou 2018c, p. 3), mais le moteur de recherche *Solr* (https://solr.apache.org/ [Consulté le 16 juillet 2022]) qui a été utilisé pour l'indexation des données

Les auteur-e-s mettent en garde que même si cette Preuve de concept est prometteuse, elle ne prenait pas en compte la longue phase de réglage (*tuning*) qu'il va falloir réaliser pour paramétrer les différents systèmes de fouilles de données, et des *benchmarks* qu'il faudra créer pour améliorer les différents systèmes de classification (Gaudinat, Knafou 2018c, p. 24). Pour valider les fonctionnalités de fouille de données proposées, elles ont également été soumises à un panel d'expert-e-s (*L5.3*).

Dans le Rapport final (*Axe 2*, *L6*), les auteur-e-s définissent en détail l'architecture fonctionnelle qu'ils recommandent pour *ArchiSelect*, comprenant un schéma fonctionnel global (Figure 15), de l'organisation et de la hiérarchie des éléments autour du document et l'héritage des métadonnées, des étapes d'extraction, analyse et indexation et de l'organisation des métadonnées et de cycles d'analyse (Figure 5).

Figure 5 : Cycle d'analyse pour enrichir les données



(Gaudinat, Knafou 2018a, p. 13)

En effet, le concept prévoit qu'une fois indexé par le moteur de recherche, il sera possible d'appliquer sur tous les documents de l'index de nouvelles analyses pour l'enrichir en nouvelles métadonnées.

Le rapport final présente également une maquette d'exploration pour la fouille de données (Annexe 4) qui propose une exploration de lots de données via un paradigme de recherche type moteur de recherche avancé. Enfin, le rapport fournit quelques recommandations techniques sur le choix du moteur de recherche et une planification d'implémentation et conclut

sur des spécifications fonctionnelles en sélectionnant neuf fonctionnalités « cruciales »[24] et « prioritaires »[25].

3.2.1.3 Axe 3 : Interface utilisateur

L'axe 3 est une étude concernant les fonctionnalités générales et l'interface utilisateur, cette partie a été réalisée en collaboration de l'agence *User experience Telono* basée à Genève et est constitué du livrable *ArchiSelect Persona & Scénarios* (Egger, Desjobert, Bongiovanni 2018) ainsi qu'une maquette en ligne. Les auteur-e-s ont ainsi créé deux *personas*, des archétypes d'utilisateurs – une archiviste évaluatrice et un archiviste validateur, et pris l'exemple de trois scenarios pour l'utilisation fictive d'*ArchiSelect*, un exemple est reproduit en Annexe 5.

3.2.2 Projets internationaux

3.2.2.1 *InterPares*

Un groupe de chercheuses et de chercheurs se réunit dès 1994 autour de Luciana Duranti, basée à la *University of British Columbia*, puis prend forme dès 1999 sous le nom *d'InterPares*. L'équipe s'intéresse d'abord à ce qu'est un document numérique en utilisant les principes de la diplomatique. Leur diagnostic est que la principale différence se situe dans la *fiabilité* et l'*authenticité* du documents, ce qui appelle des procédures claires pour préserver l'intégrité numérique et le maintien du lien archivistique (*archival bond*[26]) (Rajotte 2010, p. 76).

Le projet est ensuite organisé en « phases », la phase 1 (1999-2001) *The Long-Term Preservation of Authentic Electronic Records*, visait avant tout à développer des connaissances théoriques et méthodologiques pour la préservation à long terme et mettant en place des groupes de travail (*Task forces*) dont une sur l'évaluation (Eastwood et al. 2005) constitué dans le but de définir si les méthodologies d'évaluation pouvaient être valides dans le contexte numérique (Rajotte 2010, p. 77) *;* phase 2 (2002-2007) *Experiential, Interactive, Dynamic Records*, voir Duranti et Chabin (2004) et Duranti (2007) ; phase 3 (2007-2012) *Theoretical Elaborations into Archival Management (TEAM)*. Et la phase 4 (2012-2019) *InterPares Trust* (*ITrust)* qui se focalise sur les données en ligne.

Plus récemment, le projet *InterPares Trust AI* (*ITrustAI*) (2021-2026) explore les possibilités de l'utilisation de l'intelligence artificielle en archivistique, détermine les risques et les bénéfices de l'AI, s'assure que les concepts et principes archivistiques façonnent le développement d'une AI responsable et valide l'application de l'AI responsable en proposant des études de cas et démonstrations (InterPARES Trust AI 2021a). À titre d'exemple, le projet propose dans le cadre de l'étude *AI Tutorials (GS01)* [27] des tutoriels pour l'utilisation d'outils

[24] Fonctionnalité d'importation sous forme de système de fichiers ; extraction de contenus textes et métadonnées ; extraction et indexation de documents ; statistique (comptage d'éléments discrets et continue) ; recherche ; filtre ; interfaçage des services de fouilles ; fonctionnalité de spécification le versement ou l'élimination des lots de données ; rapport synthétique. (Gaudinat, Knafou 2018a, p. 19)

[25] Importation de plans d'archivage provenant d'outil d'évaluation prospective ; importation d'un export d'un Système documentaire ; Importation métadonnées de dossiers d'activité ; fonctionnalité pour parcourir répertoires pour lister fichiers, répertoires, extensions et taille ; calculer les métriques archivistiques ; créer métriques *combo* ; gestion des métriques *combo* ; gestion *mapping* entre métriques de données et métriques archivistiques ; voir documents dans le temps ; extraction de mots récurrents. (Gaudinat, Knafou 2018a, p. 19-20)

[26] À ce propos voir *The Archival Bond* par Luciana Duranti (1997).

[27] https://github.com/UBC-NLP/itrustai-tutorials [Consulté le 19 juin 2022]

de NLP et ML. Dans le même cadre, Ken Thibodeau dirige une étude *ITrust AI Metastudy (MA06)* dès juin 2022 qui répertorie l'utilisation des différentes méthodes et produits de l'AI en lien avec l'archivistique (InterPARES Trust AI 2021b).

3.2.2.2 *Using AI for digital selection in government* (The National Archives UK)

The National Archives UK (TNA) mènent une recherche explorant les possibilités offertes par les outils impliquant l'AI. Les TNA ont ainsi mandatés cinq fournisseurs d'outils propriétaires pour qu'ils utilisent leurs outils de classification automatique sur un jeu de données fourni par les TNA : *Adlib Elevate*, *Amazon Web Services*, *Microsoft Azure*, *InSight* par *Iron Mountain* et *Records 365* par *RecordPoint*.

Les principaux résultats sont condensés dans le rapport *Using AI for Digital Records Selection in Government: Guidance for records managers based on an evaluation of current marketplace solutions* (2021a). Si la conclusion reste que l'AI ne remplacera pas l'archiviste, les résultats sont tout de même prometteurs, les outils peuvent être impliqués dans les tâches d'évaluation au sein de collections de fichiers structurés ou semi-structurés. Les principaux enseignements de l'étude sont :

- une préparation minutieuse des données améliore significativement le résultat ;
- les solutions les plus sophistiquées n'offrent pas forcément les meilleurs résultats, ainsi l'adéquation de l'ensemble des fonctionnalités et la compatibilité avec l'environnement technologique doivent être pris en compte parallèlement à la performance brute ;
- les archivistes auront besoin d'une formation technique et l'accès à l'expertise de la *data science* pour déployer ces outils avec succès. (The National Archives UK 2021a, p. 4)

Le projet fournit également un rapport préliminaire sur l'étude de marché des outils (The National Archives UK 2020), un article sur leur *Benchamrking Tool* (Venkata 2020) ainsi que les rapports des fournisseurs, disponible sur le site du projet[28]. Par ailleurs, les TNA mène le projet *Plugged In, Powered Up*[29] de conservation numérique qui met à disposition ressources, études de cas, et *workflows*.

3.2.2.3 *BitCurator Project*

Le *BitCurator Project*[30] a comme ambition de rendre accessible les outils de *digital forensics* aux institutions patrimoniales (Lee et al. 2012) en proposant des outils *open source*. Le projet propose *BitCurator Environment*[31] qui permet la création de *forensic disk images* ; l'analyse de *filesystems* ; l'extraction de métadonnées ; et l'identification de données personnelles (BitCurator s.d.). *BitCurator Access*[32], offre des outils d'assistance aux institution

[28] https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/research-collaboration/using-ai-for-digital-selection-in-government/ [Consulté le 9 août 2022]
[29] https://www.nationalarchives.gov.uk/archives-sector/projects-and-programmes/plugged-in-powered-up/ [Consulté le 9 août 2022]
[30] Issu d'une collaboration entre la *School of Information and Library Science (SILS), University of North Carolina* et le *Maryland Institute for Technology in the Humanities (MITH), University of Maryland* et financé par la *Andrew W. Mellon Foundation*. Les outils sont maintenus par le *BitCurator Consortium*. https://bitcurator.net/ [Consulté le 4 août 2022]
[31] https://github.com/BitCurator/bitcurator-distro/wiki/Releases [Consulté le 4 août 2022]
[32] https://github.com/BitCurator/bitcurator-access/wiki [Consulté le 4 août 2022]

patrimoniales: « *in both redacting and providing access to data from disk images* » (Lee 2018, p. 2722).

Entre 2016 et 2018 le projet *BitCurator NLP*[33] développe notamment des outils de *Topic modeling* (voir 5.3.2.4.1). Pour l'extraction de texte et autres tâches de NLP, *BitCurator NLP* intègre toute une série d'outils *open source* (Hutchinson 2020, p. 164). Les *topic models* sont visualisé par *pyLDAvis*[34] (voir Sievert et Shirley (2014)). Goodman (2019) fait une très bonne description des fonctionnalités que propose l'outil de *topic modeling*, *bitcurator-nlp-gentm* et ses applications dans l'évaluation archivistique. Hutchinson est cependant plus critique sur les fonctionnalités de reconnaissance d'entités nommées (NER) proposées par *BitCurator* qu'il juge assez limitée (2020, p. 164).

3.2.2.4 AIMS

Le projet AIMS[35] (*An Inter-Institutional Model for Stewardship*) a été mené entre 2009 et 2011 et visait à identifier une méthodologie de gestion d'archives numériques, le projet propose ainsi un cadre (*framework*) sous forme d'un « livre blanc » *AIMS Born-Digital Collections : An Inter-Institutional Model for Stewardship* (AIMS Work Group 2012) proposant une approche pratique de type *best practice*. Les différentes étapes y sont décrites et documentés voire modélisées en *Workflow,* renvoyant vers des logiciels. Bien que très exhaustif, le document date déjà d'une décennie et n'aborde l'évaluation uniquement sous sa forme technique (voir Annexe G du *livre blanc*).

3.2.2.5 *Archivematica* (*Artefactual*)

Archivematica[36] est la solution *open source* de préservation numérique compatible OAIS d'*Artefactual*, le développeur d'*AtoM*. L'outil dispose également d'une *Appraisal tab*[37], certes rudimentaire, mais comprenant une fonctionnalité de tag, de reclassement, description et visualisation. L'outil n'inclut momentanément pas encore de *Natural language processing*, mais serait compatible, soit directement soit connecté par des *workflows* (Hutchinson 2020, p. 12).

Pour des retours d'expériences voir les articles de Michael Shallcross de la *Bentley Historical Library* (2015; 2016a; 2016b; Shallcross, Deromedi 2012) et l'étude de cas de Mumma et al. (2011)

3.2.2.6 *Community Owned digital Preservation Tool Registry* (COPTR)

COPTR[38] n'est pas un projet qui implémente des solutions d'archivage numérique, mais un registre ou plateforme visant à aiguiller l'utilisateur vers des outils de préservation numérique à long terme. Les outils sont ainsi classés en fonction de l'étape du cycle de vie (séquencé en *Create or Receive* ; *Ingest* ; *Preservation Planning* ; *Preservation Action ; Access, Use and Reuse* ; *Store* ; *Dispose* ; ou *Cross-Livecycle Functions*) dans lequel il se situe et les

[33] https://github.com/BitCurator/bitcurator-nlp/wiki [Consulté le 4 août 2022]
[34] https://pyldavis.readthedocs.io/en/latest/ [Consulté le 4 août 2022]
[35] https://dcs.library.virginia.edu/aims/white-paper/ [Consulté le 15 août 2022]
[36] https://www.archivematica.org/fr/ [Consulté le 9 août 2022]
[37] https://www.archivematica.org/fr/docs/archivematica-1.13/user-manual/appraisal/appraisal/#analysis-pane [Consulté le 9 août 2022]
[38] https://coptr.digipres.org/index.php/Tools_Grid [Consulté le 9 août 2022]

fonctionnalités qu'il propose. COPTR permet d'ailleurs également aux institutions d'y déposer et mettre à disposition leur *workflows*[39] de *digital curation*.

Il existe d'autres initiatives inspirés *par* ou collaborant *avec* COPTR, tel le *wiki* de la *Deutsche Nationalbibliothek*[40] ou encore le projet du CECO présenté ci-dessous (voir 3.2.3). Fait intéressant, si ces plateformes proposent énormément d'outils pour diverses étapes de la gestion d'archives numériques, l'évaluation, si elle n'est pas simplement laissée de côté, renvoie principalement vers des outils de déduplication et de visualisation d'arborescence, révélatrice d'une impuissance face à la problématique.

3.2.2.7 *OSSArcFlow*

OSSArcFlow[41]*,* pour *open source software archival workflow* est un projet investiguant des possibilités de *workflows* de préservation numérique intégrant des outils *open source* dont trois grandes plateforme de logiciels libres *BitCurator*, *Archivematica* et *ArchivesSpace*. Le but étant que les *workflows* générés puissent être généralisés aux environnements utilisant d'autres plateformes et applications. Une quinzaine d'institutions étasuniennes, principalement des universités ont ainsi modélisé leur *workflow* en indiquant à quelles étapes les outils *open source* sont été utilisés (disponible sur le site indiqué sous *Project Outputs – Born Digital Archiving Workflows, 2018*). Mais là encore, l'évaluation est peu traitée.

3.2.2.8 Autres projets et logiciels

Le logiciel *open source ePADD*[42] est spécialisé dans le traitement archivistique de courriels avec une application d'évaluation intégrant du NLP, tout comme *RATOM*[43] également spécialisé sur l'évaluation de courriels.

Vitam[44] (Valeurs immatérielles transmises aux archives pour mémoire) est un programme interministériel de l'État français visant à développer une solution logicielle libre d'archivage numérique qui ne propose cependant à notre connaissance aucune fonctionnalité pour l'évaluation archivistique.

Citons encore le projet *Paradigm* des Universités d'Oxford et Manchester, certes plus d'actualité (2005-2007), mais qui eut le mérite de proposer un *Workbook*[45] pour les archives privées en abordant la question de l'évaluation.

3.2.3 Projets en Suisse

Les projets suisses d'envergure autour de l'archivage numérique avec un focus spécifiquement sur l'évaluation sont à notre connaissance très rares. Dans le domaine de l'archivage numérique à long terme, nous pouvons citer le projet *Gal@tae* dans le canton de Genève (Dunant Gonzenbach, Ducry 2014), le système de *records management* ECM basé sur la solution *OpenText* en Valais. Également en Valais, citons le projet en collaboration avec Basma Makhlouf Shabou qui propose un *modèle de maturité de l'évaluation.* Enfin l'outil *Struc-*

[39] https://coptr.digipres.org/index.php/Browse_COW_workflows [Consulté le 9 août 2022]
[40] https://wiki.dnb.de/pages/viewpage.action?pageId=134715087 [Consulté le 12 août 2022]
[41] https://educopia.org/ossarcflow/ [Consulté le 9 août 2022]
[42] https://library.stanford.edu/projects/epadd [Consulté le 25 juin 2022]
[43] https://ratom.web.unc.edu/ [Consulté le 16 juin 2022]
[44] http://www.programmevitam.fr/ [Consulté le 16 juin 2022]
[45] https://wayback.archive-it.org/org-467/20161101115525/http://www.paradigm.ac.uk/index.html [Consulté le 16 juin 2022]

tool (anciennement *OS-Tool*) proposé par les Archives fédérales suisses (AFS) permet la mise en place de calendriers de conservation mais possède également un onglet « évaluation », la valeur archivistique du document étant attribué tant par l'office versant que par les AFS (Makhlouf Shabou 2015b, p. 202).

Le Centre de coordination pour l'archivage à long terme de documents électroniques (CECO) mène et accompagne des projets qui visent à fournir des instruments de base autour de l'archivage électronique et contribuent à la normalisation en étant notamment actif dans l'association *eCH*. À notre connaissance le CECO s'est jusque-là peu concentré sur l'évaluation archivistique spécifiquement. Cependant, le récent projet « 20.039 Collections de fichiers »[46] propose un *workflow* pour l'archivage numérique de collections de fichiers en examinant les outils et interfaces disponible et compile une série de *best practices*, semblable à ce que propose le projet AIMS. Le CECO, pour donner suite à ce projet, propose une série de rencontres pour échanger autour d'exemples et cas concrets proposés par les participants, les résultats sont ensuite mis à dispositions sur le wiki du projet[47].

[46] https://kost-ceco.ch/cms/20-039-collections-de-fichiers.html [Consulté le 21 juin 2022]
[47] https://www.kost-ceco.ch/kostwiki/doku.php [Consulté le 3 août 2022], uniquement disponible en allemand au moment de la rédaction.

4. Le contexte documentaire neuchâtelois

Afin de replacer notre travail dans son contexte, il convient de présenter le contexte institutionnel, légal et documentaire dans lequel se trouve le canton de Neuchâtel. Ce contexte et plus spécifiquement le concept *AENeas* a déjà fait l'objet d'un travail de Master, effectué par Nils Veuve (2021), également archiviste aux Archives de l'État de Neuchâtel (AEN), nous nous appuierons notamment sur ce travail dans cette partie.

4.1 Les Archives de l'État de Neuchâtel (AEN)

L'Office[48] des Archives de l'État de Neuchâtel est affilié au Département de la justice, de la sécurité et de la culture (DESC) et fait partie du Service de la culture (SCNE) au même titre que l'Office du patrimoine et de l'archéologie (OPAN) et du Laténium, musée archéologique cantonal. Les AEN comptent en 2021 6.3 équivalents plein temps, dont l'archiviste cantonal, l'archiviste cantonal adjointe, un archiviste-informaticien, un gestionnaire d'information, quatre archivistes, une secrétaire ainsi que deux apprenti-e-s. Les archivistes sont réparti-e-s en deux pôles : le *front-office*, responsable de la salle de lecture, du travail d'inventaire et de la mise en valeur des archives et un pôle plus axé *back-office*, appelé à l'interne « ÜL », pour *Überlieferungsbildung*, en charge de la supervision de l'administration et donc de l'évaluation archivistique.

Les AEN ont pour « tâches de superviser l'archivage dans l'administration cantonale, d'évaluer les documents d'activité lorsque leur durée d'utilité administrative et légale est échue, de constituer, conserver et communiquer les archives ayant un intérêt historique ou juridique permanent. Par ailleurs, il conseille et soutient les communes dans le domaine de l'archivage. » (Conseil d'État de Neuchâtel 2021, p. 184). Les AEN supervisent plus de 250 entités archivistiques et réceptionnent en moyenne une trentaine de versements administratifs et une quinzaine de dépôts ou dons d'archives privées en format papier par an pour un équivalent de 340 ml pour l'année 2020 (Conseil d'État de Neuchâtel 2021).

4.2 Cadre légal

Le 1er janvier 2012 entre en vigueur dans le canton de Neuchâtel une nouvelle Loi sur l'archivage (LArch) (442.20) remplaçant la Loi sur les archives de l'État du 9 octobre 1989. La LArch est complétée par le Règlement d'exécution du 29 avril 2013 (RLarch). Le cadre législatif est largement inspiré par la Loi fédérale sur l'archivage (LAr) du 26 juin 1998 (152.1).

Dans son rapport à l'appui du projet de loi, le Conseil d'État de Neuchâtel souligne que la LArch doit permettre « une prise en charge des documents tout au long de leur cycle de vie, de leur création à leur élimination ordonnée ou à leur conservation permanente sous forme d'archives définitives » (2010, p. 2), trois objectifs principaux sont énoncés, la LArch doit permettre d'« organiser l'archivage des documents » ; « prévoir l'archivage à long terme des documents électroniques » ; et d'« adapter l'accès aux archives aux principes de la société de l'information » (2010, p. 5).

Les entités organisationnelles sont responsables de gérer et conserver « d'une manière ordonnée leur documents jusqu'à l'expiration de leur délai d'utilité administrative et légale »

[48] Les Archives de l'État de Neuchâtel sont désormais organisées en « Office » et non plus en Service. Dans un souci d'harmonisation, nous continueront cependant à utiliser l'acronyme « AEN » et non « OAEN ».

(LArch, art 6, al. 1). Le principal outil pour ce faire est la mise en place d'un référentiel archivistique, soit un Plan d'archivage (PA) selon les directives des AEN (RLarch, Art. 5). De plus, les entités organisationnelles doivent nommer un « Préposé à la gestion de documents » (RLarch, Art. 6). Les AEN déterminent « la valeur archivistique des documents » (LArch, Art. 6, al. 2), et, une fois versés, sont responsables de la « conservation définitive et du classement des documents archivés » (LArch, Art. 10, al. 1).

Le cadre législatif neuchâtelois prend également en compte l'archivage numérique, les entités doivent ainsi tenir compte « des exigences de l'archivage lors de la conception ou du choix de leurs systèmes de gestion électronique des données. » (LArch, Art. 6, al. 3) et suivre les recommandations des AEN dans la mise en place de systèmes de gestion électronique des données (Art. 7, al. 1). Les AEN gèrent une « infrastructure informatique conformes aux normes professionnelles » dédié à la « conservation définitive des archives » (RLArch, Art. 15, al. 1) et « élabore un concept et des directives sur l'archivage à long terme des documents électroniques » (RLArch, Art. 15, al. 3).

4.3 *Records management* dans l'administration cantonale

Si jusque dans les années 1990 la production documentaire des services de l'administration neuchâteloise était classée selon des plans de classement structurés par département selon le principe de la gestion par affaire, cette tradition semble avoir disparue avec l'introduction de l'informatique. Depuis, de grandes disparités entre les structures de classement des différentes entités peuvent être observées. Ainsi l'environnement informatique et la gestion documentaire en général de l'administration cantonale et des entités organisationnelles soumises à la LArch se caractérisent par une grande hétérogénéité (Veuve 2021, p. 19-20).

En effet, les entités organisationnelles doivent certes respecter le cadre organisationnel imposé par l'Office de l'organisation (OORG), le cadre informatique mis en place par le Service informatique de l'entité neuchâteloise (SIEN) ainsi que les directives des AEN, mais de fait, les entités organisationnelles s'organisent librement. La conséquence en est une production documentaire extrêmement disparate : certaines entités produisant encore une part importante de leur documentation en format papier, d'autres sont passées à une gestion électroniques des affaires. C'est aux entités, si elles le souhaitent, de s'approcher du SIEN pour mettre en place des disques partagés ou une solution type Gestion électronique des documents (GED). Cependant, aucun outil de gestion électronique des affaires uniques, de type GEVER n'est en place ni est prévu de l'être. De fait, il n'y a pas dans le canton de Neuchâtel pas de réelle volonté politique à mettre en place une politique de *Records management* au sein de l'administration et encore moins une stratégie de cyberadministration (Veuve 2021).

4.4 L'archivage numérique

4.4.1 Le concept *AENeas*

L'absence d'une politique de *records management* digne de ce nom et une production documentaire pour le moins disparate au sein de l'administration neuchâteloise qui en résulte ont fortement conditionné le concept *AENeas*[49], le concept d'archivage numérique

[49] Archives de l'État de Neuchâtel *electronic archiving system* (AENeas) – *Aeneas*, Énée en français, aurait fondé la civilisation romaine après avoir fui Troie, tombée aux mains d'Ulysse

neuchâtelois. Le concept résulte d'une analyse détaillée de la LArch et vise à prendre en compte le contexte professionnel spécifique à chaque unité organisationnelle. Le concept illustre les étapes du processus d'archivage comme définies par la législation et décrit les moyens numériques qui garantissent la constitution et la conservation pérenne d'archives (Archives de l'État de Neuchâtel 2016). *AENeas* se concrétise notamment sous la forme de la *SuiteArchi*, une série d'outils couvrant les besoins pour couvrir le processus d'archivage.

Il y a dans le concept *AENeas* sans aucun doute un « parti pris », qui est celui d'être le moins intrusif possible : les outils de la *SuiteArchi* devraient ainsi altérer le moins possible le processus de gestion des affaires des entités et la réalité quotidienne des collaborateurs et ainsi être compatibles avec l'environnement de production documentaire des entités organisationnelles. Cette compatibilité passera par le développement de « connecteurs » permettant la gestion du cycle de vie et l'archivage pour l'ensemble des applications métiers utilisées (Veuve 2021, p. 21).

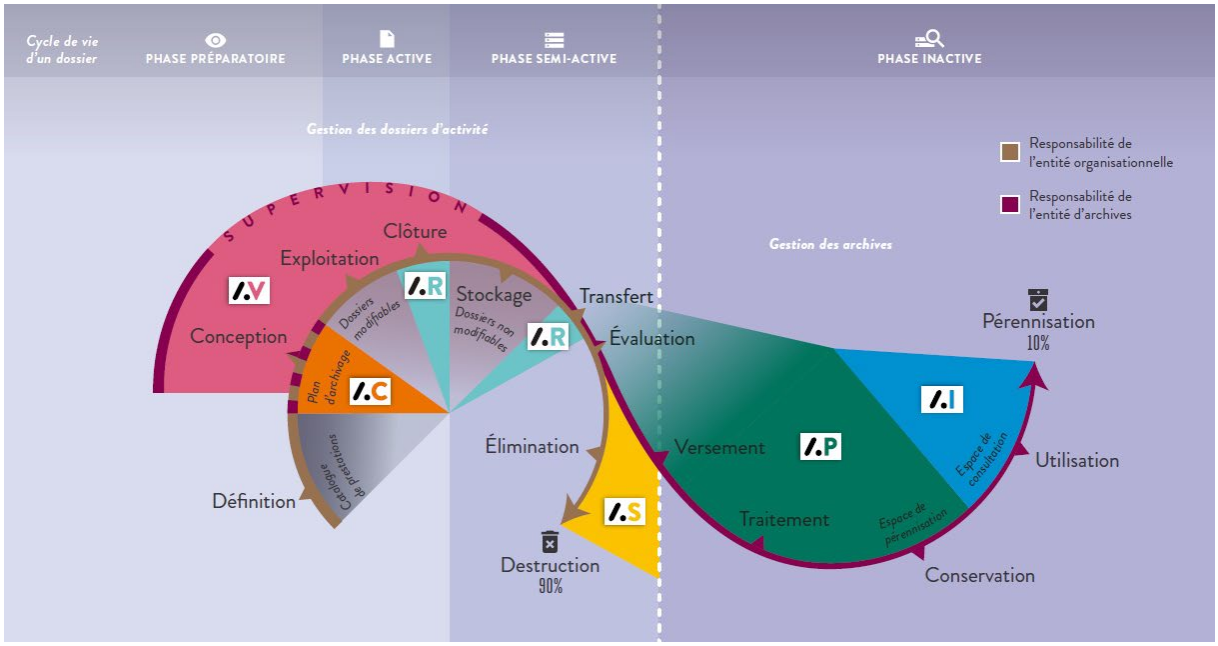
AENeas fait donc le pari d'une stratégie moins rigide qu'une approche de type GEVER, avec l'indéniable inconvénient de devoir s'adapter à la grande hétérogénéité des applications métiers. En revanche, cette approche « non intrusive » devrait permettre l'accès à l'intégralité de la production documentaire « organique » de l'administration.

4.4.2 La *SuiteArchi*

AENeas se concrétise par une série d'outils couvrant les besoins du processus d'archivage, articulés autour du cycle de vie documentaire (Figure 6). Les archivistes responsables du concept *AENeas* ont ainsi identifié le besoin de six outils qui sont soit acquis soit développés à l'interne par les AEN grâce au crédit d'investissement qui a accompagné la promulgation de la LArch. L'ordre d'acquisition ou de développement se fait dans l'ordre logique du cycle de vie, la plupart étant encore en phase de conception. De fait, seul le premier logiciel du cycle (*ArchiClass*) a été développé, le deuxième (*ArchiRef*) est, au moment de la rédaction de ce texte, sur le point de l'être. Dans la partie qui suit nous définissons la fonction et le périmètre des outils en question. La définition des besoin et le périmètre d'*ArchiSelect*, l'outil d'évaluation archivistique, sera, pour des raisons évidentes, traitée de manière plus approfondie en 4.4.3.

par sa célèbre ruse. *Aeneas* symbolise le passage d'un ancien monde (papier) vers un monde nouveau(numérique) (Archives de l'État de Neuchâtel 2016, p. 7).

Figure 6 : Schéma du cycle de vie documentaire et positionnement des outils de la *SuiteArchi*[50]



(Archives de l'État de Neuchâtel s.d. b)

4.4.2.1 *ArchiClass*

ArchiClass est l'outil qui permet la mise en place du plan d'archivage (PA), soit un plan de classement enrichi de métadonnées (du type : support, délai de clôture, durée d'utilité administrative et légale ou sort final pressenti) qui doit être validé par tous les acteurs. À terme, toutes les entités organisationnelles devront disposer d'un PA validé (RLarch, Art. 5). Défini dans le concept comme « la pierre angulaire » du processus d'archivage (Archives de l'État de Neuchâtel 2016, p. 20), l'outil permet à l'archiviste d'effectuer l'évaluation prospective et poser des bases saines au cycle de vie documentaire. Bien que prenant en compte tant la production documentaire papier que numérique, le PA peut être ensuite implémenté dans une système de serveur partagé ou de gestion électronique des documents (GED) (Veuve 2021, p. 32).

4.4.2.2 *ArchiRef*

Puisque dans le contexte informatique neuchâtelois, bon nombre d'outils informatiques tel les systèmes de fichier et autres applications métiers ne permettent pas une gestion des affaires satisfaisante, les AEN ont identifié le besoin d'un logiciel permettant de rendre effectif le passage d'une étape à l'autre du document dans son cycle de vie, dont les propriétés sont définies dans le plan d'archivage. *ArchiRef* permet ainsi de réaliser des revues de dossiers à intervalles réguliers qui permettront la déclaration de l'ensemble de la production documentaire[51] de l'entité concernée en définissant le statut de chaque dossier dans le cycle

[50] « AV » correspond à *ArchiVision*, « AC » à *ArchiClass*, « AR » à *ArchiRef*, « AS » à *ArchiSelect*, « AP » à *ArchiPeren* et « AI » à *ArchiInfo*.

[51] Le bon fonctionnement d'*ArchiRef* sera donc passablement tributaire de ses interactions avec *ArchiClass*, afin de pouvoir importer le plan d'archivage en vigueur, mais également avec l'ensemble des applications métiers identifiés dans le plan d'archivage afin de pouvoir leur envoyer un avis de clôture ou de transfert pour les dossiers concernés. Des connecteurs devront ainsi être développés entre *ArchiRef* et les applications métiers. Le grand défi sera le développement de ces connecteurs *a fortiori* l'identification des applications métiers

de vie documentaire qui fournit aux entités organisationnelles une vision d'ensemble de leur production documentaire, listant les dossiers candidats à la clôture et ceux arrivées à l'échéance de leur durée d'utilité et donc candidats à la proposition aux AEN (Veuve 2021, p. 33)

4.4.2.3 *ArchiVision*

ArchiVision sera l'outil de suivi des entités organisationnelles, le *cockpit* à partir duquel les archivistes piloteront la *SuiteArchi*. *ArchiVision* fournira aux archivistes des AEN les informations utiles sur les entités organisationnelles : informations générales de contexte, activités de l'entité, préposé, plan d'archivage et l'historique des versions, applications métiers concernées et décisions d'évaluations passées.

4.4.2.4 *ArchiPeren*

La préservation à long terme sera prise en charge par *ArchiPeren*, qui s'inscrira dans le modèle OAIS, correspondant aux exigences de la norme ISO 14721 (2012). Le modèle OAIS identifie quatre fonctionnalités principales : le stockage, la gestion des données, l'administration et la planification de la pérennisation. L'outil importe les dossiers clos ou lots de données à préserver. Il importe aussi les requêtes des utilisateurs qui souhaitent y accéder. Il exporte le résultat des requêtes sous la forme de paquets, appelé DIP (Archives de l'État de Neuchâtel 2016, p. 25).

4.4.2.5 *ArchiInfo*

L'outil de pérennisation va devoir fonctionner avec l'outil de mise en valeur, *ArchiInfo* dont la finalité est l'accès et la communication des archives au public. Actuellement les AEN utilisent le logiciel propriétaire *Flora*[52], cet outil est cependant limité est devra *a priori* être complété voire remplacé pour garantir l'accessibilité des archives aux utilisateurs au sens des Chap. 5 « Accès des archives » et 6 « utilisation des archives » de la LArch. *ArchiInfo* sera la salle de lecture virtuelle des AEN et matérialisera l'entité fonctionnelle « Accès » du modèle OAIS (ISO 14721 :2012) (Veuve 2021).

4.4.3 ArchiSelect : définition des besoins et périmètre

L'évaluation archivistique est l'intervention majeure de l'archiviste, parfois considérée comme la plus importante voire la plus noble des fonctions (Couture 1999, p. 103). *ArchiSelect* couvre donc bien l'étape charnière entre l'entité organisationnelle et l'unité d'archivage. La LArch oblige l'entité organisationnelle à proposer à l'unité d'archives (*i.e.* les AEN) par le biais d'un bordereau de versement ou d'élimination, ses dossiers clos dont les DU sont échues (Archives de l'État de Neuchâtel 2016, p. 24).

4.4.3.1 Compatibilité à différents niveaux de maturité de gestion documentaire

L'évaluation doit se faire de manière prospective dans un premier temps à l'aide des plans d'archivages conçus dans *ArchiClass.* Dans le cas de figure où une évaluation prospective a

nécessitant ce développement. De fait, le concept *AENeas* ne pourra prendre en compte l'ensemble des applications métiers, mais *ArchiRef* permettra la prise en compte de l'environnement numérique des entités organisationnelles et la captation d'un taux élevé de la production documentaire organique pour son archivage (Veuve 2021, p. 33-34).

[52] https://flora.decalog.net/ [Consulté le 21 juillet 2022]

eu lieu, *ArchiSelect* servira avant tout à la confirmation des sorts finaux pressentis et permettra, si ce n'est une automatisation, une systématisation du processus.

Cependant, cette forme d'évaluation n'est, dans le concept *AENeas*, pas suffisante et ce pour plusieurs raisons. En premier lieu, les administrations publiques utilisent l'informatique au quotidien depuis deux à trois décennies et si nous observons une tendance à prévoir des systèmes réfléchis qui suivent les recommandations archivistiques, la documentation plus ancienne, elle, est souvent déposée de manière informe sur les serveurs ou autres supports (Oguey, Schneiter 2018).

L'évaluation « rétrospective » devra donc venir à bout d'un gigantesque passif peu ou pas structuré, un passif « en vrac » encore amené à croître, pour la simple raison que l'élaboration de plans d'archivages pour chaque entité organisationnelle prendra des années. D'autre part, tant qu'il n'y aura pas de volonté d'harmonisation et la mise en place d'une politique de *records management*, le vrac continuera à s'accumuler.

Nous pensons par ailleurs que ce constat est également du moins partiellement valable pour les contextes où une cyberadministration et une gestion électronique des documents de type GEVER a été mise en place. En effet, Stephan Lenartz observe que :

> *[…] malgré les dispositions légales, l'uniformisation de l'organisation du travail administratif n'a pas été atteinte, ce manque d'uniformisation n'a cependant pas été compensé par l'accroissement de la coopération entre les services d'archives et les autorités dans le cadre du passage au numérique. Nous devons donc nous attendre à ce que les services d'archives soient de plus en plus confrontés à l'évaluation de collections de fichiers non structurés issues de contextes administratifs et institutionnels, malgré les efforts d'introduire une gestion électronique par dossiers.*[53] *(2020, p. 17)*

À ces cas de figure vient encore s'ajouter la prise en charge d'archives privées (LArch, art. 10, al. 3) où l'évaluation prospective est, par principe, impossible.

4.4.3.2 Sort final pressenti

Mais plus fondamentalement, sans remettre en cause l'importance de l'évaluation prospective et son potentiel d'automatisation, les archivistes des AEN ont la conviction qu'on « ne saurait se passer du regard humain » et que l'archiviste doit rester « maître de l'évaluation » (Oguey, Schneiter 2018). Cette idée se matérialise dans la notion de *sort final pressenti*[54], qui permet à l'archiviste de modifier le sort final défini lors de l'évaluation prospective, par exemple en cas de développements ultérieurs de thématiques sociétales importantes ou débats politiques qui rendent le sort final attribué initialement plus pertinent (Veuve 2021; Dunant Gonzenbach 2022). Cette logique se retrouve dans les Art. 7 et 8 de la LArch qui obligent les entités à continuer à proposer l'ensemble de leurs dossiers arrivés en fin de DU pour qu'ils soient

[53] Traduction libre de l'auteur : « […] *trotz aller rechtlichen Vorschriften keine vollkommene Vereinheitlichung der bürokratischen Arbeitsorganisation erreicht wird, was bisher auch nicht von den im Rahmen der Digitalisierung eingeleiteten Bemühungen um eine intensivierte Zusammenarbeit von Archiven und Behörden aufgefangen wurde. Daher ist damit zu rechnen, dass Archive in Zukunft, trotz der Bemühungen um die geregelte Aktenführung, in größerem Umfang mit Dateisammlungen aus behördlichen oder institutionellen Kontexten konfrontiert sein werden.* » (Lenartz 2020, p. 17)

[54] À ce propos, voir le billet de blog d'Anouk Dunant Gonzenbach *Le sort final pressenti : une tentative pragmatique d'application de l'archivistique post-moderniste ?* (2022).

soumis à l'évaluation archivistique qui confirmera ou infirmera le sort final pressenti (Veuve 2021, p. 48).

Nous l'avons vu, *ArchiSelect* doit donc être en mesure de prendre en charge des propositions d'archives de niveau de maturité de gestion documentaire (et donc de structure, de format et de volumétrie) très diverses. Les fonctionnalités d'aide à la décision vont ainsi passablement varier selon la nature de la proposition. Si nous intégrerons dans notre réflexion l'intégralité de ces scenarios, l'emphase sera mise sur le cas de figure le plus complexe, soit l'évaluation de propositions d'un niveau de maturité moindre, voire du « vrac » numérique.

4.4.3.3 Fonctionnalités

Nous l'avons vu dans notre état des lieux (Chapitre 3), l'évaluation est encore la grande absente tant des *workflow* de traitement d'archives numériques que des « boîtes à outils » mis à disposition aux archivistes, trop souvent abordée partiellement voire mise de côté. Certes, il existe des outils assemblés en *workflow* qui soutiennent l'évaluation, mais ils sont trop nombreux, ont une courbe d'apprentissage trop élevée et nécessitent une formation spécialisée (Belovari 2017, p. 56). *ArchiSelect* devra ainsi faire l'objet d'un développement sur mesure.

L'outil va devoir prendre en charge des données d'une grande hétérogénéité de taille, format, structure et de niveau de maturité de gestion documentaire. Il devra permettre l'extraction de métadonnées et exploiter celles-ci pour proposer des possibilités d'une visualisation facilitée, pour une meilleure appréhension des données à évaluer. Les termes d'« aide à la décision » et « bras de levier » sont également évoqués dans la documentation (Archives de l'État de Neuchâtel 2016; Oguey, Schneiter 2018; Makhlouf Shabou, Tièche 2018a; Veuve 2021) et lors des discussions avec le mandant[55].

Une base de travail misant sur le *data mining* a été posée par un mandat donné à la HEG-GE, ce même mandat propose également une première réflexion autour des fonctionnalités qui viendront soutenir l'archiviste dans sa tâche d'évaluation, une réflexion que nous poursuivrons dans ce travail en nous basant sur une analyse des pratiques d'évaluation analogique et des développements récents dans le domaine.

[55] Entretien avec le « Pôle évaluation » des AEN, 11 juillet 2022, aux Archives de l'État de Neuchâtel.

5. Résultats et analyse

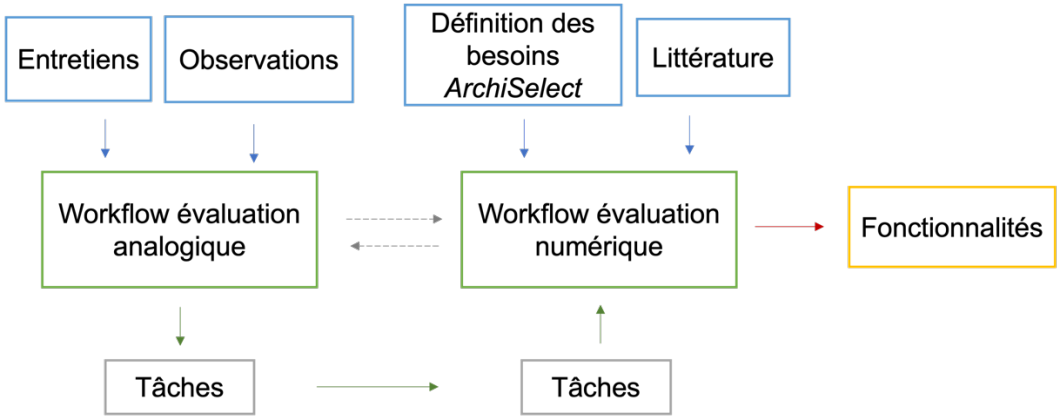
5.1 État des lieux de l'évaluation archivistique aux AEN

Pour rappel, le premier des trois objectifs de ce travail est de restituer et analyser les pratiques d'évaluation archivistique analogique ayant cours aux sein des Archives de l'État de Neuchâtel. Pour cela, nous avons appliqué un *modèle de maturité de l'évaluation*[56] pour analyser politiques, processus, méthodes et critères actuelles. Cette étape nous a permis de gagner une vision précise de l'existant. Puis nous avons prélevé et commenté le *workflow* de l'évaluation analogique en place (voir 5.1.1) en nous basant sur nos propres observations et les entretiens menés[57].

Dans la cadre de ce travail, un cadre stratégique pour l'évaluation archivistique aux AEN a également été réalisé (voir 5.1.2), dont les principes guideront les choix opérationnels dans le développement du logiciel *ArchiSelect*.

Dans une deuxième partie (voir 5.2) nous avons assemblé les éléments observés en un *workflow* pour l'évaluation de documents numériques, afin de pouvoir ensuite faire des propositions de fonctionnalités pertinentes venant soutenir ces tâches (partie 5.3) et susceptibles d'être implémentées dans l'outil *ArchiSelect*.

Figure 7 : Démarche méthodologique



5.1.1 *Workflow* de l'évaluation de documents analogiques actuel

5.1.1.1 Démarche

Pour le prélèvement et la modélisation du *workflow* d'évaluation analogique actuel aux AEN nous nous sommes basés sur nos observations personnelles durant notre activité professionnelle « d'évaluateur » au sein des AEN ainsi que sur les données d'entretien.

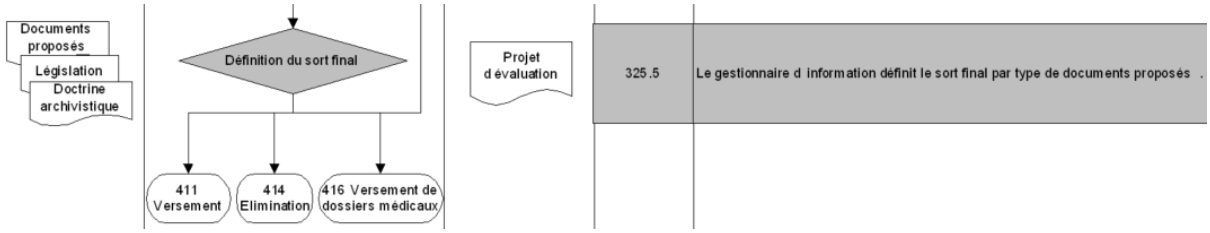
Le *workflow* a été représenté par un *Business Process Model* (BPM)[58]. Après avoir modélisé le processus de « Proposition » (Annexe 6), qui s'est avéré d'une granularité trop grossière. En effet, le périmètre d'*ArchiSelect* couvre principalement la « définition du sort final ». Dans

[56] Nous remercions au passage Basma Makhlouf Shabou de l'avoir mis à disposition.
[57] Entretiens avec le « Pôle évaluation » des AEN, 23 mai 2022 et 11 juillet 2022, aux Archives de l'État de Neuchâtel. Entretien avec Grégoire Oguey, gestionnaire d'information AEN, le 16 mai 2022, Neuchâtel.
[58] Réalisé avec l'outil *Visual Paradigm* https://online.visual-paradigm.com/fr/ [Consulté le 12 juillet 2022]

une deuxième approche nous avons donc procédé à une analyse séquentielle des différentes tâches de l'évaluation et plus précisément les tâches menant vers la définition du sort final.

Ce procédé mérite quelques précisions. Hormis le document interne « Processus SCI 325 « Proposition et évaluation définitive » » qui s'arrête, de fait, au niveau de granularité de la « définition du sort final » (Figure 8), il n'existe aux AEN aucun document – guide, manuel, *vade-mecum*, *check-list* confondus – encadrant l'évaluation archivistique au niveau cantonal.

Figure 8 : Extrait SCI « Définition du sort final »



(Archives de l'État de Neuchâtel 2019)

Nous sommes donc partis d'observations de terrain de notre propre expérience d'évaluateur aux AEN afin de créer une liste de tâches réalisées lors de l'évaluation archivistiques et les avons assemblés en un *workflow*. Si nous sommes de l'avis que trop de *workflows* d'évaluation archivistique consultés n'explicitent pas les démarches et cheminements qui mènent à l'application du sort final, prenant soins d'éviter ce moment crucial, notre démarche de rester le plus concret possible en listant de manière littérale les tâches effectuées pose également certains problèmes.

En effet, le passage du concret à un certain niveau d'abstraction, inhérent à la démarche de modélisation, est délicat et la tentation de « faire coller la réalité à la théorie » bien réelle. Ainsi l'étape de l'« Évaluation par dossier » (voir 5.1.1.2.5), soit l'application des critères, est difficilement modélisable par le *Business Process Model* et c'est de manière partiellement subjective que dans un souci de lisibilité du diagramme, nous avons regroupés les différentes tâches en sous-ensembles ensuite nommées (*unicité, exploitabilité*, *valeur historique*, *significativité des documents*) par des concepts qui n'émanent pas naturellement de nos observations.

Ce choix de regroupement par sous-ensembles de critères est d'ailleurs empreint de subjectivité également dans la mesure où, dans la pratique, tous les critères ne sont pas forcément appliqués (ou « mesurés ») et leur application ne se fait en aucun cas dans un ordre donné, mais bien souvent simultanément ou de manière diffuse : la définition du sort final se faisant souvent sur la base d'un « ressenti » général, le fameux *archivalisches Fingerspitzengefühl*[59] (Briel 2001 ; Mellifluo 2008 ; Treffeisen 2000 ; Ziwes 2020).

Précisons enfin que le but premier de cette démarche n'était pas de créer un outil ou guide d'évaluation pour les AEN, mais bien de se prêter à une *expérience de pensée* afin de pouvoir

[59] Difficilement traduisible, *Fingerspitzengefühl* signifie littéralement la « sensation que l'on a au bout des doigts », soit « doigté », au sens de savoir-faire (Leo 2022) voire « intuition ». Dans le même ordre d'idée Yves Perrotin évoque l'archiviste « flairant les dossiers comme des melons et déclarant doctement, le plus souvent sans savoir pourquoi : "Ceci est intéressant. Cela ne l'est pas !". » (Pérotin 1965, p. 140)

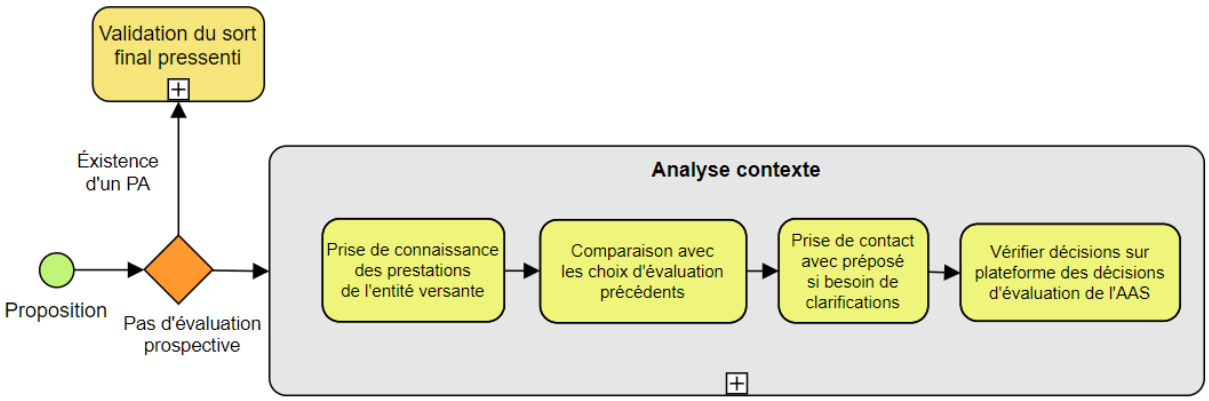
séquencer en tâches l'évaluation archivistique afin de les traduire par la suite dans un *workflow* pour l'évaluation de documents numériques.

5.1.1.2 Explicitation des étapes et tâches

Dans cette partie, nous commenterons plus en détails les différentes étapes du *workflow* en explicitant les tâches qui nous paraissent non-équivoques. Le *workflow* en son entièreté se trouve en Annexe 7. Nous avons modélisé le *workflow* de la manière la plus complète possible, soit le cas de figure d'une proposition de la part d'une entité sans évaluation prospective. Certaines étapes ou tâches deviennent ainsi vaines selon le degré de maturité de gestion documentaires ou l'origine privée de la proposition, nous le préciserons le cas échéant pour les étapes et tâches en question.

5.1.1.2.1 Analyse du contexte

Figure 9 : Analyse du contexte



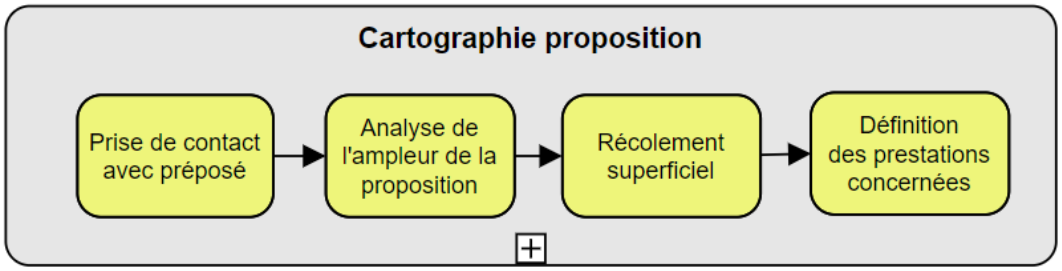
Bien que n'entrant pas dans le périmètre strict d'*ArchiSelect*, « l'analyse du contexte » fait, selon nous, partie intégrante de l'évaluation. Certes, il s'agit un processus que l'archiviste effectue en continu, mais la connaissance de l'entité versante et de l'historique des évaluations passées sont centrales pour toute évaluation archivistique. Avant toute évaluation, l'archiviste prend ainsi connaissance de ces éléments à l'aide de documents de suivi déposés sur un serveur partagé. Au besoin, l'archiviste contacte le préposé de l'entité versante pour clarifier certains points concernant la proposition.

Avant de se rendre sur le lieu des documents à évaluer, l'archiviste consulte la plateforme des décisions d'évaluation de l'Association des Archivistes Suisses (AAS)[60] afin de vérifier si d'autres services d'archives ont déjà été confrontés au même scenario.

[60] Sur le site de l'AAS, le groupe de travail « évaluation » propose une plateforme où les différents services d'archives helvétiques (fédérales, cantonales et communales, mais également privées) peuvent, à l'aide d'un formulaire, partager sur la plateforme *Bewertungsentscheide und -konzepte* soit les décisions et concepts d'évaluation. Bien que pas assez utilisé par les cantons romands, c'est une source d'information précieuse pour guider les choix d'évaluation. Au moment de la rédaction le portail était disponible uniquement en allemand https://vsa-aas.ch/ressourcen/bewertung/bewertungsentscheide-und-konzepte/ [Consulté le 5 juillet 2022]

5.1.1.2.2 Cartographie de la proposition

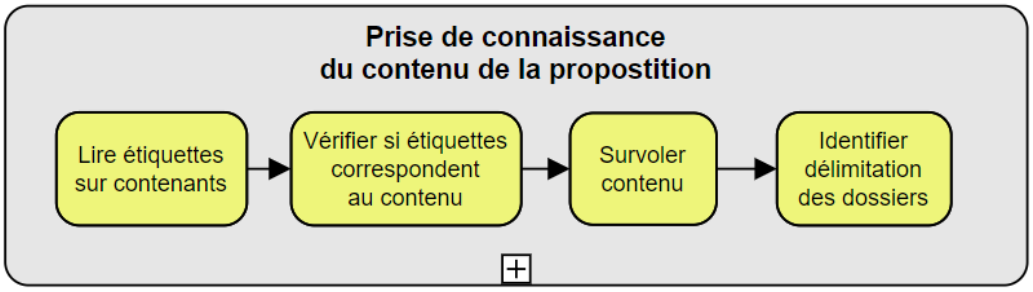
Figure 10 : Cartographie de la proposition



Cette étape a pour principal objectif de délimiter intellectuellement (et physiquement) la proposition, donnant ainsi à l'archiviste les premières indications sur l'ampleur et l'état physique des documents proposés, ce qui va permettre de planifier les ressources nécessaires à l'évaluation. En général, lors d'une mission d'évaluation, l'archiviste est accueilli par la personne préposée de l'entité versante qui lui fournit des informations sur la proposition en question.

5.1.1.2.3 Prise de connaissance du contenu de la proposition

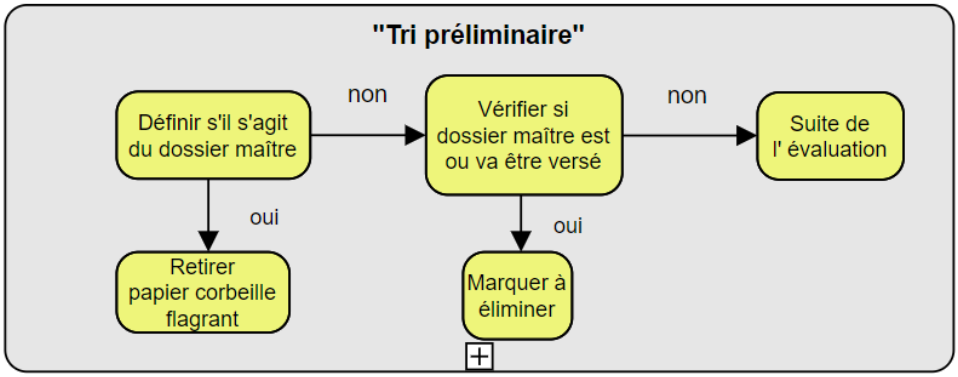
Figure 11 : Prise de connaissance du contenu de la proposition



La prise de connaissance, notamment par la lecture des étiquettes ou autres indications présentes sur les contenants ainsi que le « survol du contenu », peut paraître triviale ou anecdotique, pourtant il s'agit justement là d'une étape dont la traduction dans le numérique sera extrêmement complexe de par les différences de structures que nous avons déjà évoquées.

5.1.1.2.4 Tri préliminaire

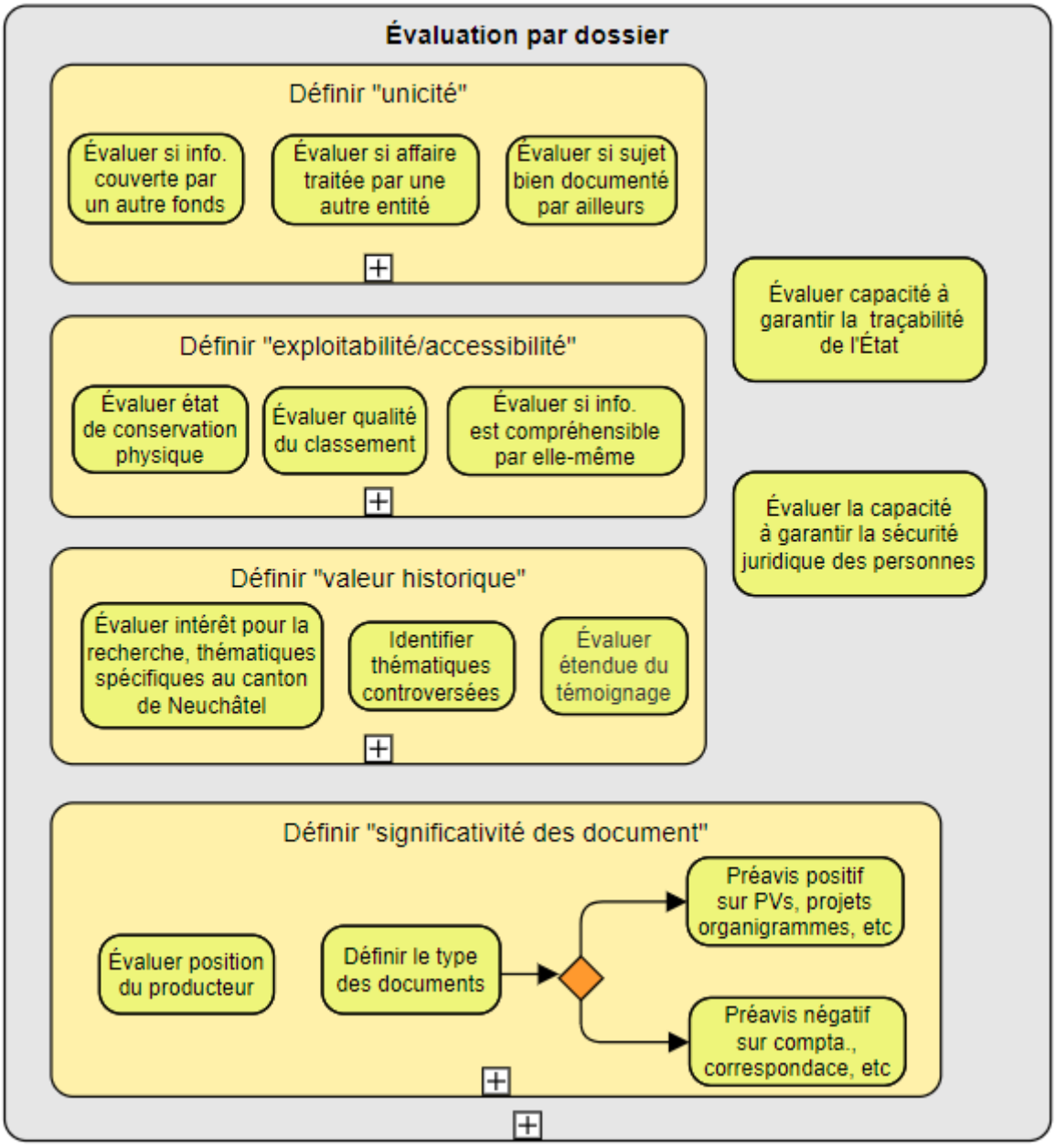
Figure 12 : Tri préliminaire



La notion de « tri » a été retenu de manière volontaire, en effet, lors de cette étape sont écartés des documents de papier corbeille, bien que cette tâche devrait être réalisée par l'entité versante, et l'archiviste ne s'y prête uniquement si l'identification est flagrante. D'autre part l'archiviste définit si les dossiers en question sont bien les dossiers maîtres et doivent être versés par l'entité en question.

5.1.1.2.5 Évaluation par dossier

Figure 13 : Évaluation par dossier



Il s'agit du cœur de l'évaluation, soit du moment où l'archiviste définit la valeur archivistique des documents proposés. Nous comprenons ici l'*unicité* comme un critère de rareté. Concrètement, lors de nos observations, les archivistes évaluent : si l'information est couverte par un autre fonds ; si l'affaire est traitée par une autre entité ; si le sujet est bien documenté par ailleurs (autres fonds, des AEN ou ailleurs).

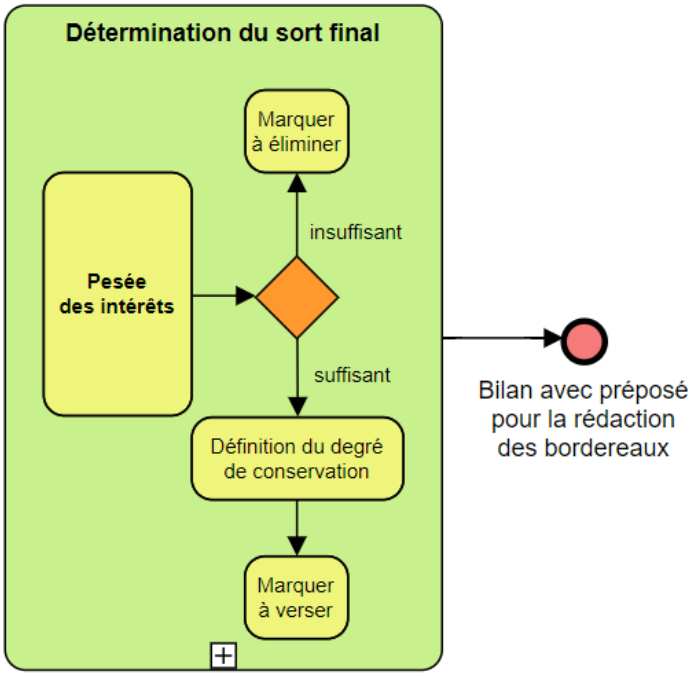
Le choix du terme *valeur historique* s'est fait faute de mieux. En effet, nos observations nous ont montrées qu'au sein des AEN, une grande importance est accordée au potentiel qu'a le document pour l'historien-ne futur, ce qui s'approche de la dimension de *preuve historique* proposée par Basma Makhouf Shabou dans son cadre conceptuel, qui correspond « à l'importance du document dans sa fonction de témoin d'un contexte historique de la société ou de l'organisation à l'origine de sa création. » (2013, p. 25), sans pour autant avoir une connotation de « preuve ». Ce que nous entendons ici par *valeur historique* s'apparente de fait plus à la dimension de *représentativité* dans le sens de « Capacité des archives à permettre un témoignage significatif, riche et exhaustif des différents éléments du contexte organisationnel de leur création. » (Makhlouf Shabou 2012a, p. 132).

Enfin, dans le processus d'évaluation, les archivistes des AEN sont attentifs à la capacité qu'ont les documents à garantir la traçabilité de l'État et les droits des personnes, inscrits dans la LArch (Art. 2, al 1 et 2).

Nous pouvons donc observer une présence moins marquée de critères se rapportant aux dimensions de *preuve crédible* (authenticité et fiabilité) et en moindre mesure d'*exploitabilité* (notamment *juridique*), des éléments sur lesquels nous reviendrons plus en détail (voir 5.3.1).

5.1.1.2.6 Détermination du sort final

Figure 14 : Détermination du sort final



L'archiviste évalue ainsi pour chaque dossier ces différentes dimensions, pesant les intérêts puis marque chaque dossier d'un « E » pour les dossiers à éliminer et « V » pour les dossiers à verser, c'est la détermination du sort final. Une fois tous les dossiers de la proposition ainsi marqués, l'archiviste fait un bilan avec le ou la préposé-e pour que celui-ci puisse reconditionner les dossiers et rédiger les bordereaux de versement et élimination.

5.1.1.3 Proposition de recommandations pour l'optimisation du *workflow* d'évaluation analogique

Sans prétention de pouvoir facilement optimiser les pratiques d'évaluation actuelles, nous ferons tout de même ici quelques modestes recommandations, dont certaines ont d'ailleurs déjà été discutées aux AEN durant la réalisation de ce Mémoire.

Si dans l'ensemble les pratiques d'évaluation nous semblent parfaitement cohérentes, l'aspect de l'unité de la doctrine appliquée pourrait être améliorée, non sur le fond – selon nos observations une doctrine commune est bien appliquée – mais sur la forme. Si les choix d'évaluation et les cas problématiques sont bien discutés en binôme voire en séance du pôle d'évaluation, il n'existe en revanche que peu de suivi de ces échanges. En effet, une capitalisation de ce savoir peut être mis en place de manière plus systématique à l'aide de rapports d'évaluation synthétiques.

D'autre part, un certain nombre de principes directeurs pourrait être énoncés – les bases d'un cadre stratégie dans le chapitre suivant est un bon début – et des pistes d'application sous forme d'un guide d'évaluation ou *vade-mecum* réalisé*,* afin de s'assurer que les pratiques d'évaluation ne soient pas dépendantes des personnes actuellement actives au sein des AEN.

Enfin, l'analyse du *workflow* de documents analogique et les entretiens menés[61] révèlent une prépondérance de la dimension de *représentativité* dans les critères d'évaluation, voire une quasi-absence de la prise en compte des indicateurs d'*authenticité* et de *fiabilité*. Cette question sera discutée dans la partie 5.3.1.2, cependant nous ne pouvons que recommander qu'une réflexion interne soit menée à ce sujet, à plus forte raison dans la perspective de l'évaluation et l'archivage numérique.

5.1.2 Cadre stratégique

Aucun document formalisé ne définissant la politique d'évaluation au sein des AEN au moment de la rédaction de ce travail, il a été défini dans les sous-objectif de rédiger, dans le cadre de ce travail, une proposition qui prendra la forme d'une liste de principes, une vision commune, qui servira de cadre stratégique pour l'évaluation et guidera les choix opérationnels dans le développement du logiciel *ArchiSelect*.

Cette liste a été créé lors d'un atelier commun sous forme de *brainstorm*[62] avec l'équipe responsable de l'évaluation :

[61] Entretiens avec le « Pôle d'évaluation » des AEN, 23 mai 2022 et 11 juillet 2022, aux Archives de l'État de Neuchâtel.

[62] La question « Que vous évoque "l'évaluation archivistique" dans le cadre de la constitution de la création d'un cadre stratégique pour l'évaluation ? » a été posée aux cinq participants (l'archiviste cantonal, le gestionnaire d'information et trois archivistes dont l'auteur) et les réponses sous forme de mots-clés ont été récoltés à l'aide d'un générateur de nuages de mots (Les mots-clés récoltés sont : Cadre normatif ; choix ; cœur du métier ; conservation ; compréhension entité ; élimination ; essentiel ; fonction centrale ; mémoire collective ; miroir sociétal ; modestie ; numérique ; objectivité ; oubli ; périmètre d'application ; pragmatisme ; pratique ; qualité ; réflexion ; responsabilité sociale ; sélection ; subjectivité ; toute-puissance ; transparence ; tri ; unicité de doctrine ; valeur patrimoniale ; versement ; vision historique.) Une fois le nuage réalisé il a été projeté et les mots-clés discutés un par un, certains ont été « écartés » lors de la discussion, d'autres ont été ajoutés. L'auteur a ensuite structuré ces thèmes autour d'une liste de principes qui a été soumis aux participants.

- L'évaluation est une fonction archivistique centrale et essentielle, elle doit se situer au cœur du métier d'archiviste ;
- L'évaluation doit se faire dans une perspective historique en prenant en compte la valeur patrimoniale du document. L'évaluation sert ainsi à créer une mémoire commune pour la société dans son ensemble. Le résultat de l'évaluation doit permettre de créer un miroir de la société, cependant la constitution d'une mémoire passe également par l'oubli qui doit être une part assumée de l'évaluation ;
- L'évaluation doit se faire dans une perspective de sécurité juridique de l'État et des personnes : elle documente les décisions importantes pour la collectivité et pour les personnes ;
- L'évaluation met l'archiviste dans une certaine position de toute-puissance, une responsabilité sociale dont il doit rester conscient ;
- L'évaluation doit se faire avec réflexivité, en étant conscient qu'elle porte en soi une part de subjectivité ;
- L'évaluation doit se faire avec du recul, en respectant une certaine lenteur et inertie des évènements et en évitant un interventionnisme trop marqué auprès des entités productrices d'information ;
- L'évaluation doit se faire dans la transparence et en respectant une unicité de la doctrine. Les choix d'évaluation doivent ainsi être documentés afin de laisser aux générations futures les éléments nécessaires à leur compréhension ;
- L'évaluation doit être guidé par un certain pragmatisme et ancré dans la pratique, prenant en compte les réalités de son périmètre d'application tout en appliquant le cadre normatif en vigueur.

5.2 Proposition d'un workflow pour l'évaluation de documents numériques

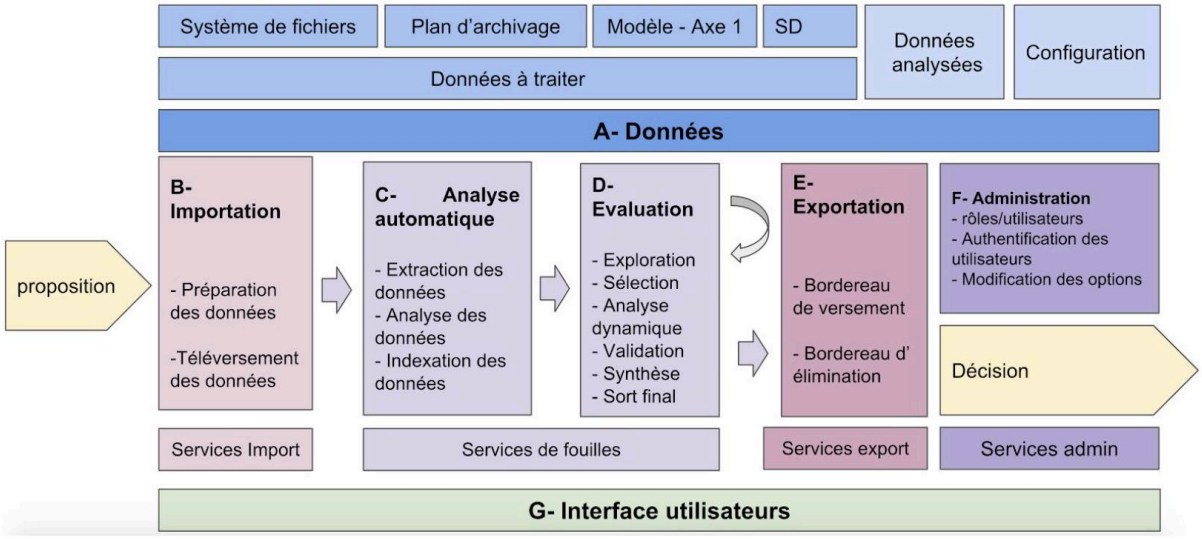
Afin de pouvoir atteindre notre troisième objectif, soit proposer des recommandations de fonctionnalités pour l'évaluation archivistique de documents numériques, nous proposons dans cette partie un *workflow* pour l'évaluation de documents numériques en nous basant sur la définition des besoins, les entretiens menés aux AEN[63], ainsi qu'en nous basant sur la littérature spécialisée (voir Figure 7). Le mandat de recherche mené par la HEG-GE propose un travail poussé sur les fonctionnalités possibles (notamment *Axe 2, L5.3, L6 et Annexe L6*) qui sera intégré à notre réflexion. L'apport de notre travail étant la prise en compte d'une analyse préalable des pratiques d'évaluation ayant cours aux AEN.

Le *workflow* proposé est ainsi compatible avec le *schéma fonctionnel global* proposé pour *ArchiSelect* (Figure 15) par le mandat de la HEG-GE. La partie A (en bleu) du schéma représente les données, la partie G (en vert) l'interface utilisateurs, tandis que dans les parties B, C, D, E et F (en violet) représentent les différentes fonctionnalités d'*ArchiSelect*. Les parties C et D, soit respectivement *Analyse automatique* et *Évaluation* étant le cœur de la fouille de

[63] Entretiens avec le « Pôle d'évaluation » des AEN, 23 mai 2022 et 11 juillet 2022, aux Archives de l'État de Neuchâtel.

données (Gaudinat, Knafou 2018f, p. 7). Notre réflexion mettra d'ailleurs l'emphase sur ces deux parties, laissant de côté les fonctionnalités d'administration.

Figure 15 : Proposition de schéma fonctionnel global pour *ArchiSelect*



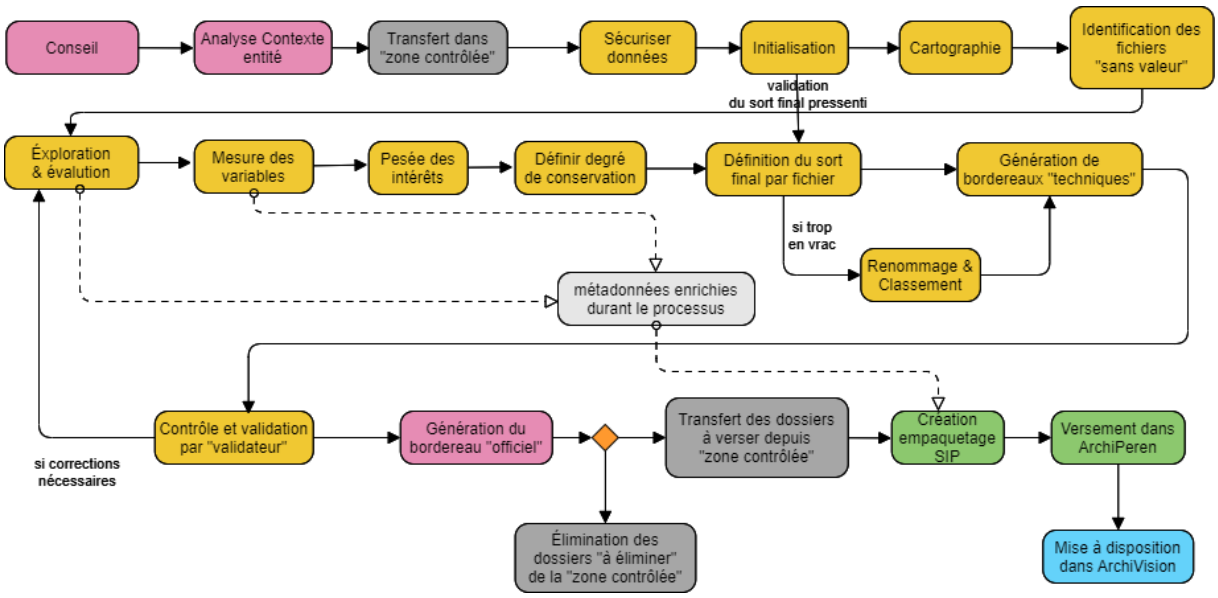
(Gaudinat, Knafou 2018f, p. 7)

Le *workflow* que nous proposons renverra à des fonctionnalités que nous jugeons nécessaires – ou pour le moins *utiles* – à la réalisation des différentes tâches de l'évaluation. Dans la partie qui suit, nous présentons la modélisation du *workflow* pour l'évaluation de documents numériques en son entier, puis spécifiquement les différentes étapes du *workflow* tout en explicitant et discutant nos choix. Chaque étape est accompagnée d'un tableau contenant les fonctionnalités proposées, le tableau entier est disponible en Annexe 8 et les définitions de certains termes techniques explicités dans le Glossaire. Certaines fonctionnalités choisies sont discutées plus en détail dans la partie 5.3. Afin d'éviter toute ambiguïté, un identifiant est attribué à chacune des fonctionnalités (f1, f2, etc.).

5.2.1.1 Workflow numérique intégral

Sur la représentation du *workflow* en son entier (Figure 16), les étapes du processus se trouvant dans le périmètre d'*ArchiSelect* sont en orange, les étapes en rose concernent *ArchiVision*, le vert *ArchiPeren* et le bleu *ArchiInfo*. Les cases grises concernent la zone protégée du transfert des données à évaluer.

Figure 16 : Proposition de *workflow* pour l'évaluation de documents numériques – vision d'ensemble



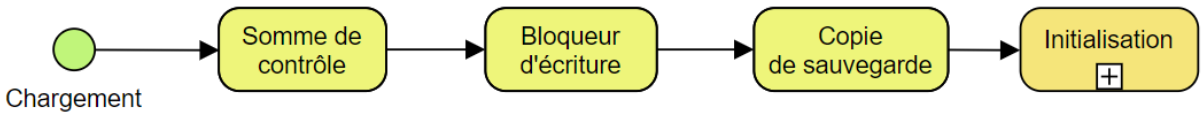
Il est prévu que les entités versantes transfèrent leurs propositions au sein d'une « zone contrôlée », la forme exacte que prendra cette plateforme est ses modalités d'utilisation ne sont pas encore définies, notamment la question de savoir si l'entité versante garde une copie ou non (en d'autres termes si elle fait un copier/coller ou un couper/coller) reste encore ouverte.

Cependant, les outils de la *SuiteArchi* sont *a priori* pensés de manière indépendante les uns des autres : *ArchiSelect* doit ainsi être capable d'analyser un jeu de données indépendamment du cheminement par lequel il y est arrivé[64]. D'ailleurs, la manière dont la proposition arrive dans la « zone contrôlée » fait partie du périmètre d'*ArchiVision* et non *ArchiSelect*.

5.2.1.2 Sécuriser les données

La première étape sera de sécuriser les données.

Figure 17 : Sécuriser les données



Bon nombre de *workflows* consultés proposent de commencer par un contrôle des virus de données proposées, cependant nous sommes de l'avis qu'il est de la responsabilité des entités versantes de faire en sorte que les données soient « propres ». Nous proposons donc de commencer par réaliser une somme de contrôle (f1) afin d'avoir une empreinte de ce qui a été versé, ce qui permettra de vérifier l'intégrité des données après chaque étape de l'évaluation.

Il est imaginable, dans le cas d'une proposition sur un disque mobile (CD, clé USB ou disque dur externe) – dans le cas d'archives privées par exemple – d'utiliser un « bloqueur d'écriture »

[64] Lors de l'entretien avec le « Pôle évaluation » des AEN du 11 juillet 2022, l'image utilisée a été qu'*ArchiSelect* analyse « ce qui se trouve dans la pièce, qu'importe comment c'est arrivé là ».

(f2) qui évite toute modification sur les données initiales, celui-ci devrait bien entendu être levé dans le cas où un classement et/ou renommage des dossiers est réalisé.

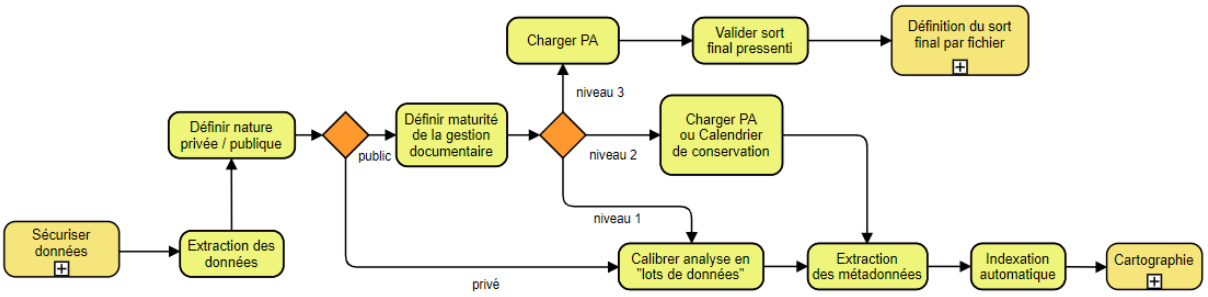
En théorie, la création d'une copie de sauvegarde (f3) pourrait être superflue dans la mesure où *ArchiSelect* ne fait pas d'*ingest* des données proposées mais analyse les données déposées dans la « zone contrôlée ». La question sera également très dépendante du fait que les entités versantes déposent une copie ou les données « originales ».

Tableau 2 : Fonctionnalités liées à la sécurisation des données

ID	Fonctionnalité	Objectif(s)	*Input*	*Output*
f1	Calcul de somme de contrôle	Création d'une empreinte de ce qui a été versé, permettant de vérifier l'intégrité des données aux différentes étapes d'évaluation	Lot de données proposées	Empreinte pour chaque fichier
f2	Bloqueur d'écriture	Empêcher les fichiers d'être modifiés ou altérés lors du processus d'évaluation	Lot de données proposées	Données bloquées
f3	Copie de sauvegarde	Création d'une copie de sauvegarde dans le cas de figure que la "zone contrôlée" subit des dommages	Lot de données proposées	Copie de sauvegarde

5.2.1.3 Initialisation

Figure 18 : Initialisation



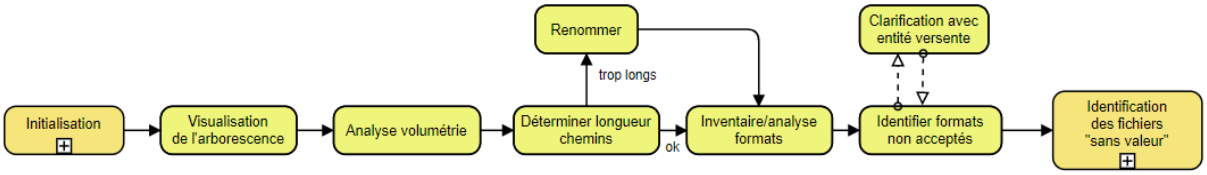
Lors de l'étape d'initialisation, Les données sont extraites, au besoin à l'aide de la reconnaissance optique de caractères (OCR) (f4), puis est défini la maturité de la gestion documentaire de la proposition (f8), qui va conditionner la suite du processus. Toute proposition issue d'un environnement documentaire inférieur à la maturité 3 (soit des entités où un Plan d'archivage validé est en place) passera par une indexation automatique des données. *ArchiSelect* importera également lors de l'initialisation les données sous forme de Systèmes de fichiers (f5) ; le Plan d'archivage (f6) ; et l'export du Système documentaire (f7).

Tableau 3 : Fonctionnalités liées à l'initialisation

ID	Fonctionnalité	Objectif(s)	*Input*	*Output*
f4	Reconnaissance optique de caractères (OCR)	Reconnaissance des documents scannés	Document scanné	Texte reconnu
f5	Importation données sous forme de SyFi	Importer des lots de données sous forme de fichiers	Lot de données proposées	Système de fichiers accessible par ArchiSelect
f6	Importation plan d'archivage	Importer le plan d'archivage, ajout des métadonnées du PA dans l'index	Plan d'archivage	Métadonnées du PA dans l'index
f7	Importation d'un export SD	Importer un export d'un SD dans l'outil 4 pour ajouter les métadonnées dans l'index	Fichier(s) ad hoc XML, JSON ou YML qui contient les métadonnées	Métadonnées de l'export d'un SD dans l'index pour chaque fichier
f8	Définition du niveau de maturité	Définir le niveau de maturité de la gestion documentaire afin de définir le niveau d'analyse	Informations sur outils d'évaluation prospective	Indication niveau 1, 2 ou 3
f9	Indexation automatique	Indexer un système de fichier dans un moteur de recherche de données structurées et de texte pour offrir des paradigmes de recherche, d'analyses et d'explorations avancées	Un système de fichiers et/ou des métadonnées extraites	Métadonnées et texte dans un index de moteur de recherche

5.2.1.4 Cartographie de la proposition

Figure 19 : Cartographie de la proposition



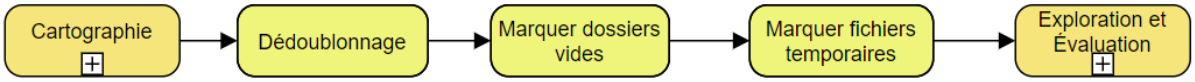
Cette étape poursuit les mêmes objectifs que la cartographie d'une proposition dans le domaine analogique (voir 5.1.1.2.2), soit appréhender la taille, la forme (format) et la structure globale de la proposition (sera discuté en 5.3.2.1). Nous proposons d'inclure dans cette étape la gestion des chemins trop longs ainsi que le marquage de formats non acceptés. En effet, si des formats non acceptés sont présents dans la proposition et que le service d'archives n'est pas en mesure de les transformer, une solution doit être trouvée en accord avec l'entité versante.

Tableau 4 : Fonctionnalités liées à la cartographie de la proposition

ID	Fonctionnalité	Objectif(s)	*Input*	*Output*
f10	Visualisation arborescence	Appréhender structure d'un lot de données	Chemins des dossiers	Cartographie de la structure
f11	Analyse volumétrie	Appréhender volumétrie d'un lot de données	Lot de données proposées	Indicateurs clés de volumétrie
f12	Inventaire formats	Appréhender formats d'un lot de données	Lot de données proposées	Liste des formats

5.2.1.5 Identification des fichiers « sans valeur »

Figure 20 : Identification des fichiers « sans valeur »



Cette étape est à mettre en parallèle avec l'étape « Tri préliminaire », elle identifie et marque les fichiers « sans valeur », où une évaluation archivistique à proprement parler n'est pas nécessaire, il s'agit de marquer « à éliminer » les doublons (f13, voir 5.3.2.2), les dossiers vides (f14) ainsi que les fichiers temporaires (f15).

Tableau 5 : Fonctionnalités liées à l'identification des fichiers « sans valeur »

ID	Fonctionnalité	Objectif(s)	*Input*	*Output*
f13	Déduplication	Supprimer doublons	Lot de données proposées	Clé de recherche identique
f14	Identification des dossiers vides	Identifier dossiers vides	Lot de données proposées	Dossiers vides marqués
f15	Identification des fichiers temporaires	Identifier fichiers temporaires	Lot de données proposées	Fichiers temporaires marqués

5.2.1.6 Exploration & évaluation

L'étape « Exploration & évaluation » est sans surprise la plus complexe, à l'instar de « l'évaluation par dossier » dans le domaine analogique, il s'agit de l'étape où l'archiviste applique les critères d'évaluation. Cette étape est appliquée uniquement pour les niveaux de maturité de gestion documentaire 1 et 2, si la proposition émane d'une entité où un PA est appliqué, le sort final pressenti sera vérifié sans, a priori, avoir recours à la fouille de données.

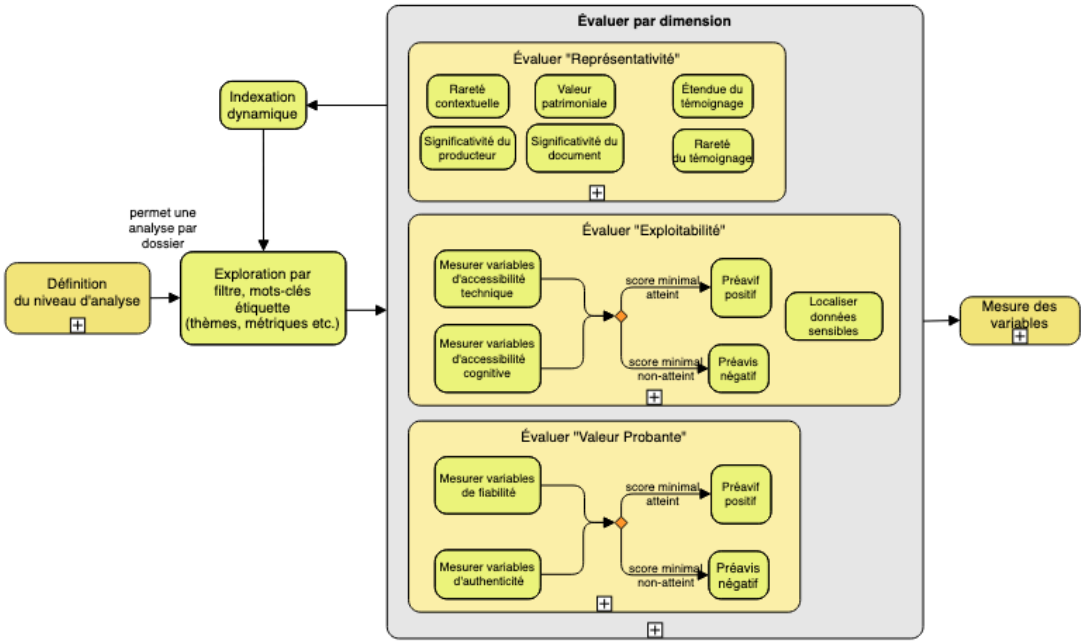
Nous pensons cependant que la maturité de gestion documentaire est à voir comme un spectre et qu'il ne sera pas toujours possible (ou même utile) de choisir entre les deux premiers niveaux, ce sera à l'archiviste de juger le « niveau d'analyse » pertinent, soit décider si la proposition est assez structurée pour pouvoir considérer certains niveaux de profondeur comme des « dossiers » ou « délimitations intellectuelles pertinentes », sur lesquelles on

pourra mesurer les différentes métriques. Ou, si au contraire la proposition est à aborder comme un « lot de données » ou « collection de fichiers non structurée » qu'il s'agira d'appréhender de manière plus horizontale, et dans quel cas, on passera vraisemblablement par une indexation par fichier. Les documents auront dans ce cas de figure un sens au sein d'une « agrégation » virtuelle créé par un système de *tag* (f27), ou alors on décidera de le structurer par un reclassement (f31).

5.2.1.6.1 Évaluation par dimension – par dossier

Si les documents proposés sont assez structurés pour pouvoir délimiter intellectuellement des ensembles, l'archiviste explore le contenu à l'aide de filtres, recherches de mots clés (f17 et f18) et par recherche d'étiquettes (par exemple par thèmes, métriques, typologie de documents issus de l'indexation automatique) tout en continuant l'indexation de manière dynamique (f20).

Figure 21 : Niveau d'analyse par dossier



Lors de ces explorations, l'archiviste évalue les différentes dimensions d'évaluation[65] (DEV) par dossier, il doit par exemple pouvoir sélectionner des éléments « par bloc » (f16), basé sur la cartographie et la visualisation de l'arborescence (f10) et les indexer et leur appliquer un sort final.

Une réflexion critique sur les dimensions d'évaluation et les variables va être menée dans la partie 5.3.1.2. Pour l'élaboration de ce *workflow*, nous sommes partis du constat que la mesure des variables de la dimension de *représentativité* – centrale dans les pratiques d'évaluation

[65] Dans ce *workflow*, pour la dimension de *représentativité*, le choix de retenir les DEV2 *rareté contextuelle, valeur patrimoniale, significativité du producteur* et *significativité du document* est basé sur leurs bons scores dans les résultats des questionnaires (voir 5.3.1.2). Tout comme *étendue du témoignage* et *rareté du témoignage*, qui bien qu'attribués dans le modèle conceptuel à la dimension de *valeur probante*, nous paraissait plus proche de la dimension de *représentativité* et semblables à ce que nous avions identifié comme « valeur historique » dans le workflow analogique.

aux AEN – n'était pas automatisable (Makhlouf Shabou, Tièche 2018b, p. 11-12 ; Makhlouf Shabou 2015b ; 2015a), les variables restant pertinentes, d'autres fonctionnalités doivent donc soutenir l'archiviste dans son analyse et son évaluation.

Nous proposons ainsi l'utilisation de fonctionnalités innovantes impliquant le *Natural Language Processing*, tel le *topic modeling* (f23), la reconnaissance d'entités nommées (f21) ou encore la classification automatique de documents (f22). Ces fonctionnalités, dont nous discuterons les possibilités d'applications plus en détail dans la partie 5.3.2.4 peuvent fournir à l'archiviste des « bras de levier » dans l'appréhension de grands volumes de données.

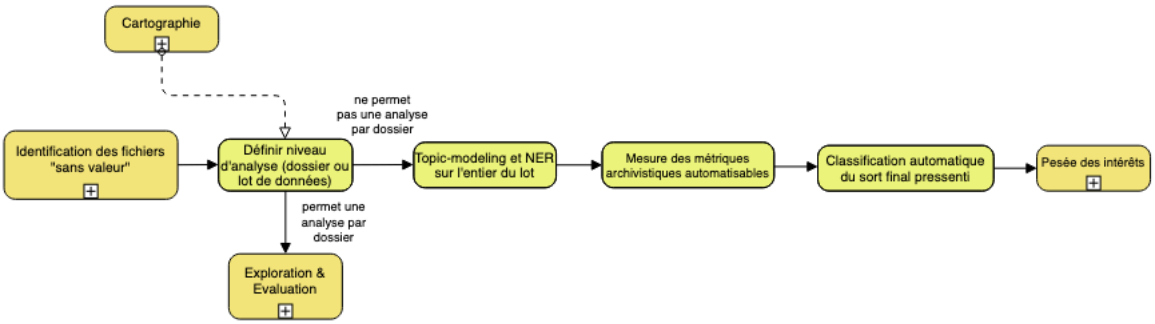
La mesure de l'*exploitabilité* et de la *valeur probante* des dossiers est quant à elle en grande partie automatisable par le mesure des métriques archivistiques (f24, voir 5.3.1). De plus, nous indiquons la possibilité de l'utilisation de la fonctionnalité de *Sensitivity review* (f28), soit la localisation de données sensibles qui sera discutée dans la partie 5.3.2.4.3.

Nous y reviendrons en conclusion, mais l'utilisation de ces fonctionnalités, spécialement celles proposées pour évaluer les différentes dimensions d'évaluation, vont évidemment devoir être testées en conditions réelles et confrontées aux réalités du terrain. La construction d'une pratique d'évaluation de données numériques qui est, pour l'heure, purement théorique aux AEN pourra certes se nourrir des possibilités techniques offertes, mais va indubitablement faire émerger d'autres besoins de fonctionnalités ou *a contrario* démontrer l'inefficacité de certaines fonctionnalités proposées ici.

5.2.1.6.2 Évaluation par « lot de données »

Si de notre point de vue l'absence *totale* de structure au sein d'un lot de données n'est possible qu'en théorie, *ArchiSelect* doit pouvoir donner des indicateurs pour des lots de données tellement en vrac qu'il faudra faire abstraction de toute arborescence.

Figure 22 : Niveau d'analyse par fichier ou « lot de données »



Dans ce « scenario du pire », l'archiviste est bien obligé de s'appuyer sur des indicateurs généraux et automatisés (ou semi-automatisés), dans quel cas nous suggérons d'utiliser les fonctionnalités de *topic modeling* (f23) afin de créer une idée générale des thématiques traitées par les documents ; d'utiliser les entités nommées présentes (f28) et la classification typologique des documents (f22).

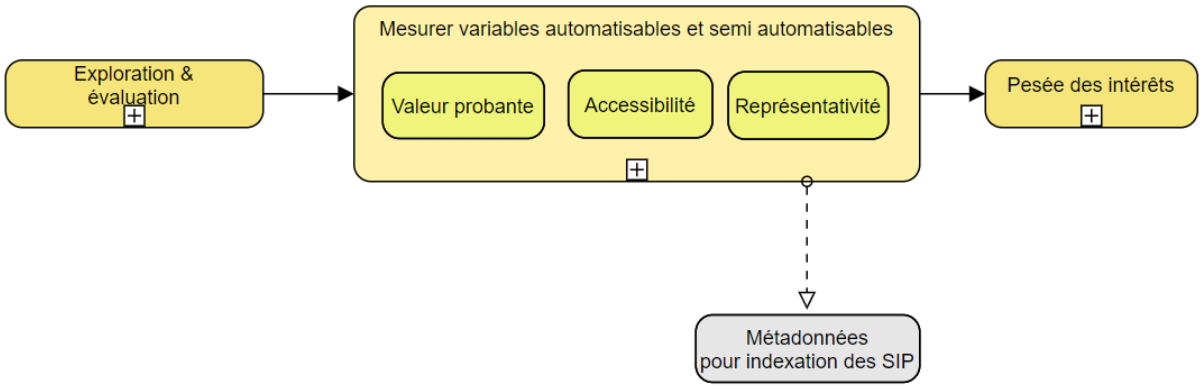
Nous avons également indiqué la possibilité de faire recours à l'utilisation de la classification automatique du sort final sur base de *machine learning* – possibilité sur laquelle nous mettons énormément de réserves pour les raisons que nous discuterons dans la partie 5.3.2.4.4.

Tableau 6 : Fonctionnalités liées à l'exploration & évaluation

ID	Fonctionnalité	Objectif(s)	*Input*	*Output*
f16	Sélection par bloc basée sur la cartographie	Pouvoir sélectionner dans la visualisation d'arborescence des blocs de données pour analyse ou indexation	Cartographie	Bloc sélectionné
f17	Recherche moteur	Chercher des documents et des métadonnées à partir de combinaisons de mots et/ou de métadonnées	Requêtes	Liste des meilleures documents pour les requêtes
f18	Recherche par facette	Filtrer les recherches avec certaines métadonnées	Requête filtre	Liste des meilleures documents pour les requêtes
f19	Prévisualisation	Pouvoir prévisualiser un fichier sélectionné	Fichier sélectionnée	Prévisualisation du fichier
f20	Indexation dynamique	Durant l'exploration pouvoir continuer l'indexation de fichiers ou blocs de fichiers	Fichier	étiquette attribuée
f21	Reconnaissance d'entités nommées (NER)	Reconnaître les entités nommées dans le texte comme noms, lieux ou institutions	Texte	Liste d'entités nommées (éventuellement position dans le texte)
f22	Classification typologique (automatique)	Classifier les documents selon une catégorie	Texte et éventuellement autres métadonnées	Liste de classe avec un indice de confiance
f23	*Topic modeling*	Appréhender rapidement les thématiques d'un corpus	Texte	Topic model
f24	Métriques archivistiques	Obtenir un "score" des DEVs de manière automatique	Métadonnées	Indicateur de confiance
f25	Visualisation document sur le temps	Voir le nombre de documents dans le temps	Documents datés et périodes de temps	Nombre de documents par période de temps
f26	Détecter cadre de classement	Détecter si un cadre de classement est en place en analysant le chemin du fichier	Système de fichier	Nom et id des dossiers d'activité
f27	*Tag*	Pouvoir marquer un certain nombre de fichiers pour garder ou "créer" le lien archivistique	Fichiers	Fichiers tagués
f28	*Sensivity review*	Identifier informations sensible dans les documents	Texte	Surlignage de données sensibles
f29	Détection signature électronique	Identifier les signatures électroniques dans les métadonnées	Fichier	Métadonnées indiquant l'existence de la signature

5.2.1.7 Mesurer variables archivistiques

Figure 23 : Mesure des variables archivistiques



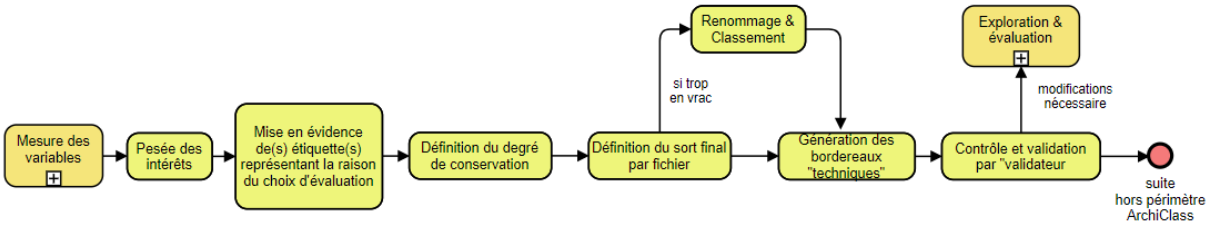
Dans cette étape nous proposons de mesurer toutes les métriques archivistiques (f24) automatisables implémentées, non plus dans un but d'évaluation des valeurs archivistiques et d'attribution d'un sort final, mais bien plus dans un but de description, afin de pouvoir fournir aux futurs utilisateurs des indicateurs sur la qualité des archives définitives. La DEV de *représentativité* est ici représentée, mais dans l'état actuel des choses, il n'est pas envisageable de la mesurer de manière automatique.

5.2.1.8 Attribution du sort final

Enfin, l'archiviste va attribuer un sort final à chaque fichier. Notre modélisation (Figure 24) représente cette étape de manière linéaire, avec une pesée des intérêts, mise en évidence des raisons du choix etc., en pratique nous imaginons cependant ce procédé comme continu et débutant déjà lors de l'exploration et parallèlement à l'indexation, l'archiviste sélectionnant

au fur et à mesure les fichiers – plus vraisemblablement des blocs de fichiers, documents ou branches de l'arborescence à verser et à éliminer.

Figure 24 : Attribution du sort final



ArchiSelect devra également disposer d'une fonctionnalité d'échantillonnage automatique (f30) ainsi que des fonctionnalités de renommage (f31) et de reclassement (f32, voir 5.3.2.3) au cas où le manque de structure est tel qu'il rend sa compréhension et la création d'un SIP impossible en l'état.

Tableau 7 : Fonctionnalités liées à l'attribution du sort final

ID	Fonctionnalité	Objectif(s)	*Input*	*Output*
f30	Échantillonnage automatique	Échantillonner de manière automatique	Liste des fichiers à échantillonner	Liste des fichier échantillonnés
f31	Renommage automatique	Si lot de données trop en vrac, permet d'imposer un système de nommage	Fichiers	Fichiers renommés
f32	Réorganiser arborescence	Si lot de données trop en vrac, permet le calssement de celui-ci	Arborescence	Arborescence réorganisée
f33	Rapport technique	Spécifier le versement ou l'élimination d'un élément du lot de données	Élément à modifier et indication de versement ou élimination (éventuellement commentaire sur la décision, date, etc.)	Élément modifié ainsi que tous les documents qui en dépendent hiérarchiquement

5.3 Recommandations fonctionnalités associées

Cette partie du travail mène une réflexion plus approfondie sur certaines fonctionnalités recommandées sélectionnées car jugées particulièrement dignes d'intérêt de par leurs possibilités techniques, mais également pour les questions techniques et de principe, qu'elles soulèvent quant à leur application. Ainsi, nous aborderons d'abord la fonctionnalité de mesure des métriques archivistiques proposée par le mandat réalisé par la HEG-GE (voir 5.3.1). Puis nous traiterons quelques fonctionnalités utiles pour les premières étapes du *workflow* avant de passer aux fonctionnalités offrant des possibilités de soutien pour la phase « Exploration et évaluation » qui impliquent l'utilisation de *machine learning*.

5.3.1 Métriques archivistiques et métriques de données

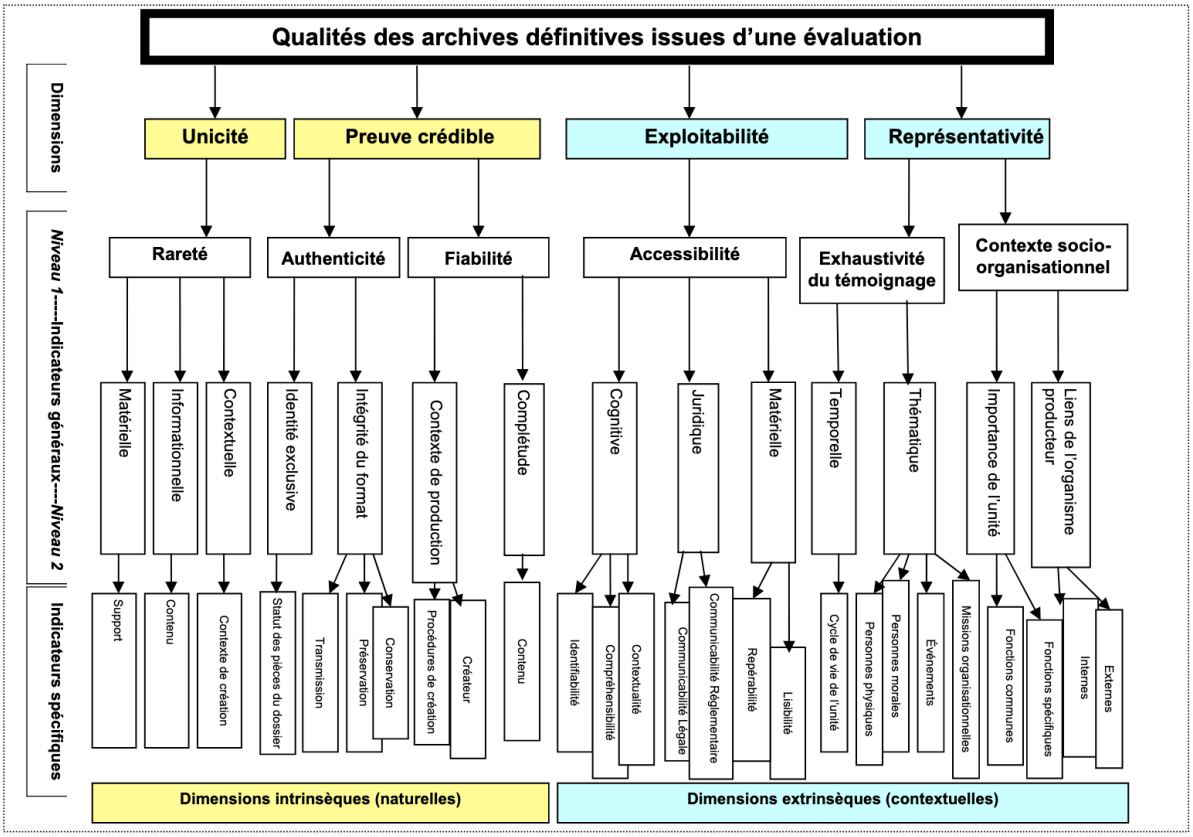
Nous considérons dans le cadre de ce mémoire les métriques archivistiques telles que proposées dans le résultat du mandat réalisé par l'équipe de la HEG-GE, comme une fonctionnalité (f24) parmi d'autres pouvant venir appuyer le travail d'évaluation de l'archiviste. Au vu de la complexité technique et des concepts théoriques qui sous-tendent ce travail, nous avons cependant décidé d'y consacrer un chapitre à part.

En effet, la preuve de concept proposée aux AEN, bien qu'accompagnée d'un rapport contenant des spécifications fonctionnelles pour le développement de l'outil, demande encore une certaine étude notamment sur le choix d'implémentation des différentes variables, comme il l'est d'ailleurs précisé dans le *Livrable 3.2* de l'Axe 1 (Makhlouf Shabou, Tièche 2018d, p. 14). Ainsi, nous proposons ici de poursuivre cette étude sur les métriques archivistiques en nous concentrant sur leur compatibilité d'application aux sein des AEN.

5.3.1.1 Origines

Pour une mise en perspective, nous pensons qu'il est éclairant de retracer le parcours bibliographique des variables issues des dimensions d'évaluation (DEV). À notre connaissance, elles émanent de la réflexion menée par Basma Makhlouf dans le cadre de sa thèse (2010) dont un résumé du même titre est publié dans la revue *Archives* (2012b), et une publication tirée sous le titre de *Comment évaluer la qualité des archives ? : Méthode et instruments de mesure des dimensions de qualité des archives définitives* (2012a). L'auteure y épluche méticuleusement les textes des principaux théoriciens de l'évaluation archivistique ainsi que la norme ISO 15489 (2001) sur le *Records management* en inventoriant toutes les qualités souhaitées pour les archives définitives. Makhlouf Shabou structure ensuite ces qualités dans un cadre conceptuel où elle distingue quatre niveaux (Figure 25) le plus général étant la dimension, comprenant l'unicité, la preuve crédible, l'exploitabilité et la représentativité.

Figure 25 : Cadre conceptuel des Qualités des archives définitives



(Makhlouf Shabou 2010, p. 127)

Dans son étude, l'auteure sélectionne deux dimensions, soit la *preuve crédible* et *l'exploitabilité* et identifie des variables pour chacune d'elle, variables qu'elle mesure ensuite sur des fonds d'archives définitives en appliquant un niveau de qualité de 1 à 5.

Basma Makhlouf Shabou, en collaboration avec Laure Mellifluo et Raphaël Rey, adapte ensuite le cadre conceptuel aux réalités du numériques dans l'étude *QADEPs : Définition et mesure des qualités des archives et documents électroniques*. En s'associant à plusieurs partenaires publics et un privé, cette étude a pour objectif de « fournir aux archivistes un outil

clair et facilement applicable pour évaluer la qualité des archives électroniques publiques dans la perspective d'une conservation pérenne dans un système de type OAIS. » (2013, p. 2).

Le cadre conceptuel ne comprend plus la dimension d'unicité, jugée moins pertinente dans le domaine numérique, tandis que la sous-dimension de « preuve historique » est rajoutée pour inclure la notion de « rareté du témoignage » (2013, p. 39). De plus, une réflexion sur le potentiel d'automatisation de la mesure des variables est menée. Toujours dans le cadre de l'étude, les variables sont mesurées sur plusieurs fonds d'archives numériques, et une évaluation de l'applicabilité et de la pertinence du cadre conceptuel réalisé. L'étude a donné lieu à un article (Makhlouf Shabou 2015a).

Enfin, comme présenté dans notre travail (voir 3.2.1), le cadre conceptuel retravaillé est proposé dans le cadre de la preuve de concept d'*ArchiSelect*, en tant que dimensions *d'évaluation* – soit comme critères pour l'évaluation archivistique, alors qu'elles étaient jusque-là utilisées comme dimension de mesure de *qualité* d'archives définitives – soit ayant déjà passé l'étape de l'évaluation archivistique[66].

5.3.1.2 Choix des métriques archivistiques

L'historique du cadre conceptuel nous enseigne notamment qu'il n'est pas figé et peut – voire doit – être adapté aux différents contextes d'application, c'est pourquoi il nous a paru utile de prolonger la réflexion entamée lors du mandat réalisé par la HEG-GE. En effet, les métriques archivistiques ont été soumises à un panel d'expert-e-s en vue de recueillir leur niveau d'accord sous forme d'une échelle de Likert, les résultats sont analysés dans le *Livrable 3.2* de l'Axe 1 (Makhlouf Shabou, Tièche 2018d) et des recommandations d'implémentation fournies.

Il nous a cependant paru pertinent de récolter la position des archivistes des AEN face aux métriques archivistiques, ceux-ci ont ainsi été discutées de manière globale lors d'un premier entretien réunissant toute l'équipe pratiquant l'évaluation archivistique[67], puis, afin de récolter le niveau d'accord plus spécifique pour chaque variable, le même questionnaire soumis aux panel d'expert-e-s par l'équipe de la HEG-GE – que nous remercions au passage pour la mise à disposition du questionnaire – dans le cadre du mandat a été soumis aux archivistes de l'AEN. Enfin, les métriques archivistiques ont été rediscutées lors d'une séance de travail commune[68] une fois que les archivistes avaient remplis le questionnaire.

L'analyse de ces données (*L3.2* de l'Axe 1, entretiens internes aux AEN, résultats des questionnaires remplis par les archivistes de l'AEN) ont fait émerger deux aspects qui, selon nous, méritent discussion et qu'il s'agira de prendre en considération lors du développement effectif d'*ArchiSelect* : d'une part une forte disparité d'approbation parmi les dimensions d'évaluation, avec la *représentativité* remportant un accord nettement plus marqué. D'autre part des possibilités d'automatisation de certaines variables mal réparties parmi les DEVs.

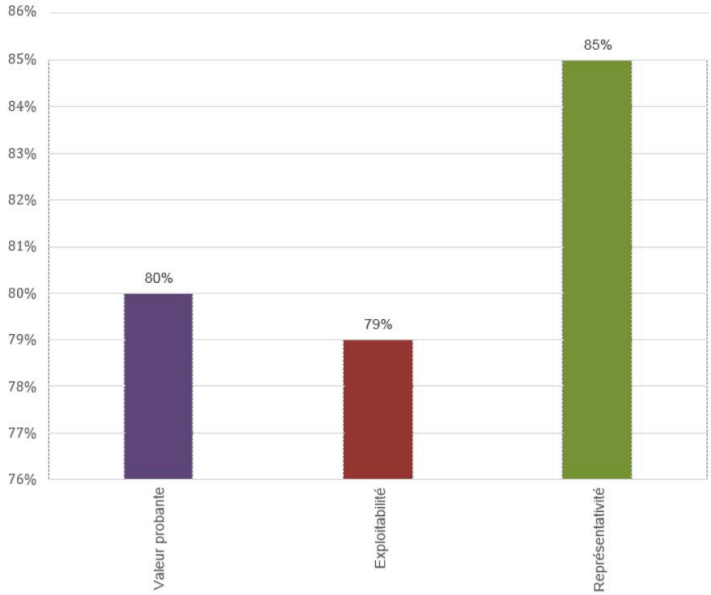
[66] Par ailleurs, la question de savoir si « mesurer les qualités des archives définitives » (QADEPs) peut être considéré *en creux* comme « évaluer la valeur d'archives intermédiaires » (DEVs) nous paraît légitime.
[67] Entretien avec le « Pôle évaluation » des AEN, 23 mai 2022, aux Archives de l'État de Neuchâtel.
[68] Entretien avec le « Pôle évaluation » des AEN, 11 mai 2022, aux Archives de l'État de Neuchâtel.

Après avoir discuté ces deux points nous proposerons un set de métriques recommandés à l'implémentation prioritaire.

5.3.1.2.1 Plébiscite de la dimension d'évaluation de représentativité

La Figure 26 démontre bien que le panel d'expert-e-s consulté dans le cadre du mandat réalisé par la HEG-GE exprime un accord bien supérieur à la *représentativité* qu'aux dimensions de valeur probante et d'exploitabilité.

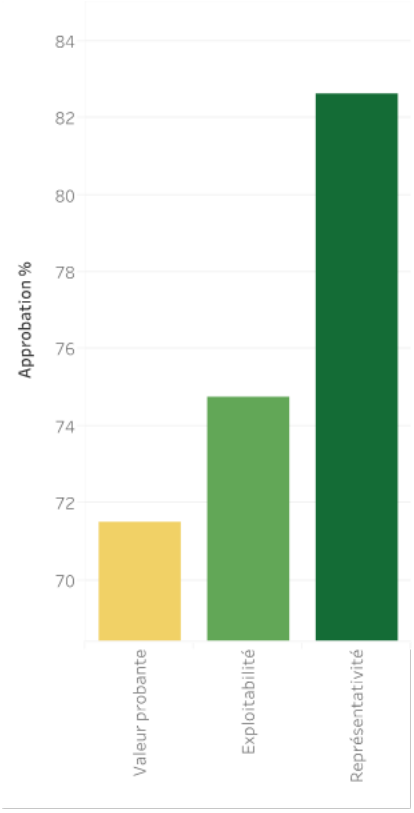
Figure 26 : Approbation par dimension d'évaluation du panel d'expert-e-s consulté par la HEG-GE



(Makhlouf Shabou, Tièche 2018d, p. 6)

Cette tendance est largement confirmée par les résultats du test réalisé auprès des archivistes des AEN (Figure 27). En effet, la différence en points est respectivement de 11 et 8 points entre la représentativité et les deux autres DEVs, contre 5 et 6 points dans le questionnaire réalisé par la HEG-GE.

Figure 27 : Approbation par dimension de la part des archivistes AEN



Par ailleurs, l'approbation de la DEV *valeur probante* est passablement tirée vers le haut par les bons scores des variables de la DEV2 *preuve historique*. Hors, si l'on se fie à la définition de cette dernière, la *preuve historique* correspond à « l'importance du document dans sa fonction de témoin d'un contexte historique de la société ou de l'organisation à l'origine de sa création » (Makhlouf Shabou 2013, p. 25) et est notamment constituée des DEV3 de *rareté du témoignage* et *étendue du témoignage*, qui pourraient, à notre avis, également être rapprochés par leur lien au témoignage à la DEV de *représentativité* définie comme la « Capacité des archives à permettre un témoignage significatif, riche et exhaustif des différents éléments du contexte organisationnel de leur création » (Makhlouf Shabou 2010, p. 123).

S'il ne s'agit pas ici d'ergoter, ces disparités entre les dimensions d'évaluation ont soulevées lors des séances de travail communes[69] de vives discussions sur l'application des DEVs dans le processus d'évaluation. En effet, dans une pure tradition latine de l'archivistique, qui épouse plus le point de vue de la mémoire collective que les intérêts juridiques et fonctionnels du producteurs (Chabin, Watel 2006, p. 115), une grande majorité des archivistes des AEN considèrent, sans mettre en doute leur validité, que les dimensions de *valeur probante* et d'*exploitabilité* ne peuvent être considérées, comme des critères d'évaluation archivistiques « durs ».

En effet, des qualités d'archives souhaitées issues de la norme ISO 15489 de *Records management*, appliquées à un vrac numérique posent une question de principe, car celui-ci aura de toute façon un très mauvais « score », cela le rend-il sans valeur archivistique pour autant ? À titre d'exemple, l'Université de Neuchâtel, qui est une entité organisationnelle soumise à la LArch, a un devoir de proposition de leurs documents arrivés en fin de durées

[69] Entretiens avec le « Pôle évaluation » des AEN, 23 mai 2022 et 11 juillet 2022, Neuchâtel.

d'utilité aux AEN de par la LArch. Imaginons le cas de figure fictif d'une collection de fichiers non structurés produite par un Professeur d'université réputé, contenant potentiellement des documents inédits non publiés proposés aux AEN : si l'on se fiait à la mesure des variables de *valeur de preuve* et d'*exploitabilité*, ces documents seraient selon toute vraisemblance jugés sans valeur archivistique. Si ce cas de figure est certes extrême, le contexte général de la gestion documentaire dans l'administration neuchâteloise, tel qu'analysé dans le Chapitre 4, ne permettra pas, à court terme d'empêcher des cas de figure semblables.

Ainsi, l'*exploitabilité*, du moins technique, semble plus considérée comme un indicateur d'*archivabilité*, immensément précieux pour l'archivage à long terme de type OAIS[70], n'entrant cependant en compte uniquement de manière secondaire dans l'évaluation archivistique (rétrospective) entendue comme « l'acte de juger des valeurs que présentent les documents d'archives (valeurs primaire et valeur secondaire » (Couture 1996, p. 3).

D'autre part, la *valeur de preuve* a été considérée lors de nos entretiens comme un *indicateur* très important, notamment à transmettre aux utilisateurs futurs, mais n'entrant pas réellement en compte lors de l'évaluation archivistique. L'exemple des « faux documents » comme source historique d'importance a été cité à plusieurs reprises, une idée par ailleurs développée par Lionel Bartolini dans la revue *Arbido* : « Alors que les normes de gestion documentaire insistent sur l'authenticité et la valeur probatoire comme caractéristiques des documents d'archives, il est bon de se souvenir que l'histoire ne s'écrit pas uniquement avec des documents originaux et que d'authentiques faux peuvent être très révélateurs d'une époque. » (2021).

Certaines variables des dimensions d'*exploitabilité* et de *valeur probante*, semblent ainsi être considérés aux AEN comme étant du ressort de la diplomatique numérique, qui ne doit pas être une « critique rétrospective d'un fonds documentaire déjà constitué en vue d'une exploitation de sources historiques […] mais bien une évaluation prospective de l'archivabilité de l'information » (Chabin 2008).

Les DEVs d'*exploitabilité* et de *valeur probante* ne doivent évidemment pas pour autant être totalement ignorées, les historiens n'étant par ailleurs pas les seuls utilisateurs des archives définitives – ils représenteraient même actuellement une minorité pour les archives numériques (O'Neill Adams 2007, p. 32). Les DEVs doivent cependant être pondérées, notamment d'après la maturité de la gestion documentaire des documents évalués, ce qui est d'ailleurs totalement compatible avec le cadre conceptuel proposé : « cet outil est modulable dans la mesure où il permet l'utilisation partielle et ciblée des indicateurs et des mesures des qualités des archives électroniques. » (Makhlouf Shabou 2013, p. 54).

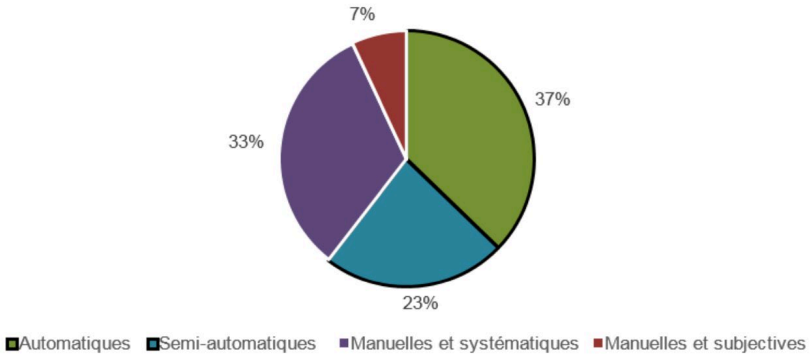
5.3.1.2.2 Critère d'automatisation

Un autre aspect ayant émergé durant l'étude du cadre conceptuel et de l'analyse de l'approbation des variables est la répartition disparate des possibilités d'automatisation. En effet, et cela a été mis en évidence dans le *Livrable 3.2* de l'axe 1, si une majorité des variables

[70] À titre d'exemple, plusieurs commentaires du type : « Pas utile dans le cadre de l'évaluation, métadonnées néanmoins utiles à conserver avec les AIP lors de conservation pérenne » ou « Option plutôt "RM", bon respect des procédures internes ou réglementaire n'impacte pas sur qualité d'archives à mon sens. » ont été fait dans le questionnaire.

disposent d'un potentiel d'automatisation élevé (Figure 28), leur répartition au sein des DEVs peut s'avérer problématique.

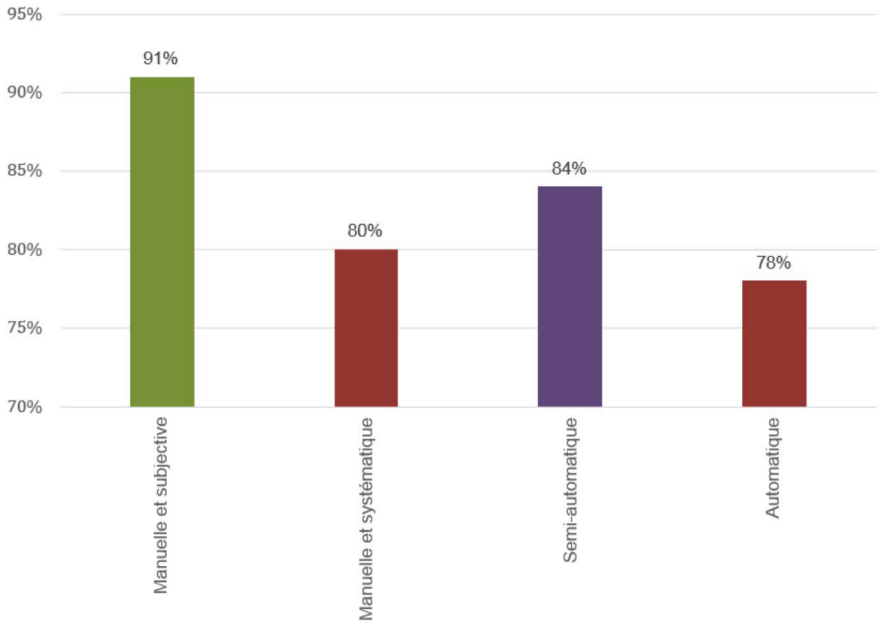
Figure 28 : Répartition des variables par critères d'automatisation



(Makhlouf Shabou, Tièche 2018d, p. 8)

Ainsi, ce sont les variables ayant reçu le plus d'approbation qui ne sont uniquement mesurables de manière « manuelle et subjective » (Figure 29).

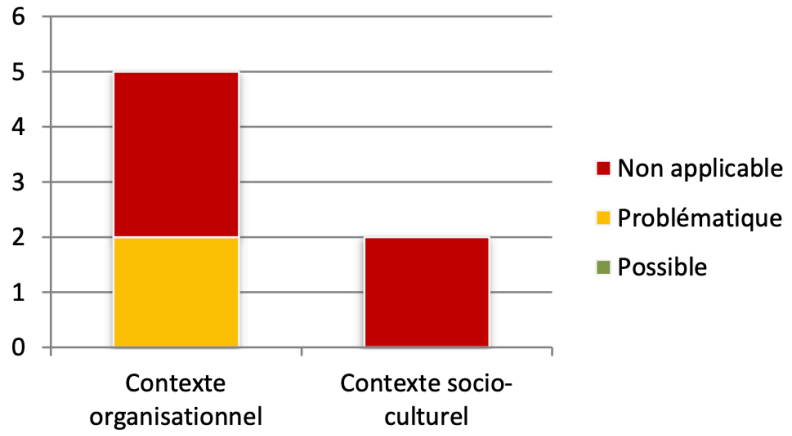
Figure 29 : Approbation par critères d'automatisation



(Makhlouf Shabou, Tièche 2018d, p. 68)

Cela est encore plus flagrant pour la DEV de *représentativité* et ses sous-dimensions de *contexte organisationnel* et *socioculturel* (Figure 30), ayant reçu la plus grande approbation et, nous l'avons démontré dans la partie précédente, est considérée dans les pratiques d'évaluation des AEN comme essentielle. « Cette situation n'est pas surprenante vu qu'il s'agit évidemment des parties où l'on s'attend à trouver le plus de subjectivité. » (Makhlouf Shabou 2013, p. 51). Reste qu'au vu de l'analyse des besoins *ArchiSelect* va devoir être capable, si ce n'est d'automatiser, d'offrir des « bras de leviers » à l'archiviste pour mesurer la *représentativité* des documents. Cet état de fait a fortement influencé le choix des fonctionnalités recommandées, notamment celles impliquant du *machine learning* (voir 5.3.2.4).

Figure 30 : Automatisation des mesures : représentativité



(Makhlouf Shabou 2013, p. 51)

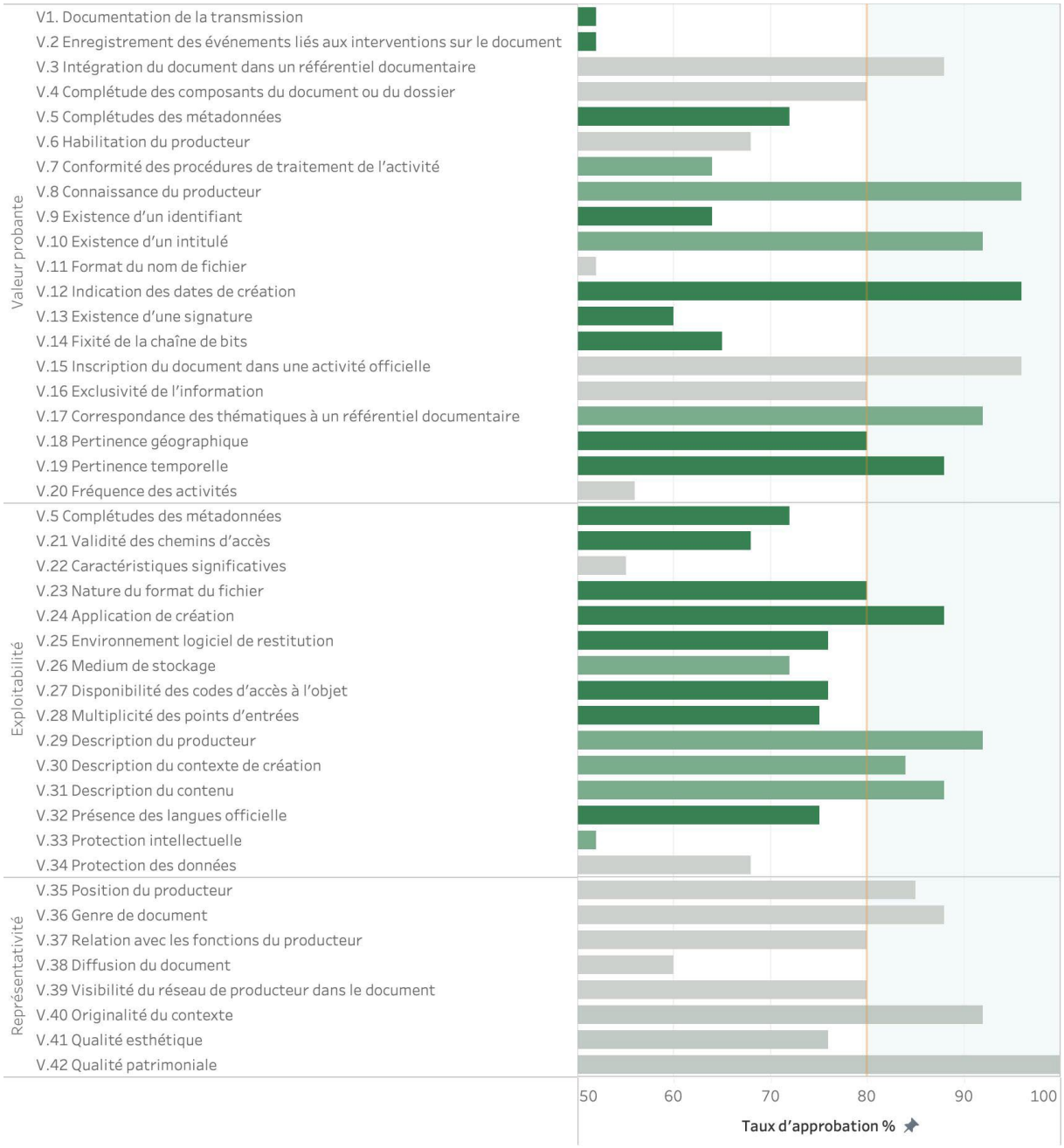
5.3.1.2.3 Set de métriques proposées à l'implémentation

Nous recommandons ci-dessous (Tableau 8) un set de métriques à implémenter en priorité. Le choix de celles-ci s'est fait sur deux critères : celui de l'automatisation, en effet, nous considérons qu'en priorité les métriques archivistiques automatisables puis semi-automatisables doivent être implémentées, car c'est bien entendu avec elles que le plus grand gain de temps et d'efficacité peut être réalisé. Le deuxième critère, subjectif dans une certaine mesure, est un taux d'approbation de plus de 80% reçu de la part des archivistes des AEN, ce qui correspond aux réponses « d'accord » et « tout à fait d'accord » (voir Figure 31).

Tableau 8 : Set de métriques proposées à l'implémentation

DEV1	DEV2	DEV3	Variables	%	Ecart-type	Automatisation
Valeur probante	Authenticité	Identité	V.8 Connaissance du producteur	96	0,45	2
			V.10 Existence d'un intitulé	92	0,89	1
	Preuve historique	Traçabilité des activités	V.15 Inscription du document dans une activité officielle	96	0,45	1
		Étendue du témoignage	V.17 Correspondance des thématiques à un référentiel documentaire	92	0,89	2
			V.18 Pertinence géographique	80	0,00	1
			V.19 Pertinence temporelle	88	0,55	1
Exploitabilité	Accessibilité technique	Lisibilité	V.23 Nature du format du fichier	80	1,22	1
			V.24 Application de création	88	1,34	1
	Accessibilité cognitive	Compréhensibilité	V.29 Description du producteur	92	0,55	2
			V.30 Description du contexte de création	84	0,45	2
			V.31 Description du contenu	88	0,89	2
			V.32 Présence des langues officielle	75	0,50	1

Figure 31 : Résultats questionnaire sur l'approbation des variables auprès des archivistes des AEN



Ceci reste bien entendu une indication. Les résultats recueillis constituent une base de travail solide, mais nous sommes de l'avis que l'approbation des variables récoltée à l'aide d'une échelle de Likert devrait se baser sur un échantillon plus vaste pour plus de représentativité, par exemple en élargissant l'échantillon aux archivistes des communes – ce qui n'a pas été possible dans le cadre de ce mémoire pour des questions de faisabilité.

D'autre part, l'écart à la moyenne élevé observée pour certaines variables, soulève la question de savoir si les testeurs disposaient d'une compréhension identique des variables. La décision définitive d'implémentation pourrait ainsi passer par une discussion des livrables au cas-par-cas, par exemple sous la forme de *focus group*. Il faudra également prendre en compte les modalités techniques et le *mapping* entre les métriques archivistiques et de fouille de données.

5.3.1.3 Avantages et désavantages d'une approche *rule-based*

L'approche des métriques archivistiques et spécifiquement du *mapping* entre métriques archivistiques et métriques de fouille de données proposée par le mandat de la HEG-GE est, à notre connaissance inédite. Innovante dans sa volonté de réconcilier métadonnées et dimensions d'évaluations, elle offre le grand avantage d'être reproductible et transparente, ce qui rend la traçabilité des décisions aisée dans la mesure où la *règle* ou le *pseudocode* appliqué peut être facilement documenté et mis à disposition dans un rapport d'évaluation. Le principe a cependant *les inconvénients de ses avantages* : en effet, la *règle* n'est par nature pas un modèle statistique mais bien un code que l'archiviste doit définir, en sélectionnant quelles métadonnées sont prises en compte, ce qui induit une part de subjectivité, qui reste ainsi inhérente à l'évaluation archivistique. Une approche hybride, mêlant règles et modèles statistiques de prévision (ML) pourrait être envisagée et testée.

D'un point de vue purement stratégique d'implémentation, il faut également souligner que le mandat, de son propre aveu, ne fournit qu'« un premier jet » (Gaudinat, Knafou 2018b, p. 13) pour une vingtaine de variables. Si une implémentation effective dans *ArchiSelect* devait être envisagée, cela représenterait encore un chantier non négligeable.

5.3.2 Fonctionnalités recommandées

Pour certaines fonctionnalités de cette partie, des exemples et illustrations de logiciels existants seront donnés à titre d'illustration uniquement et non dans le but d'intégrer lesdits logiciels à *ArchiSelect*, une potentielle intégration demanderait un examen qui dépasse les prétentions de ce travail.

5.3.2.1 Visualisation arborescence (f10)

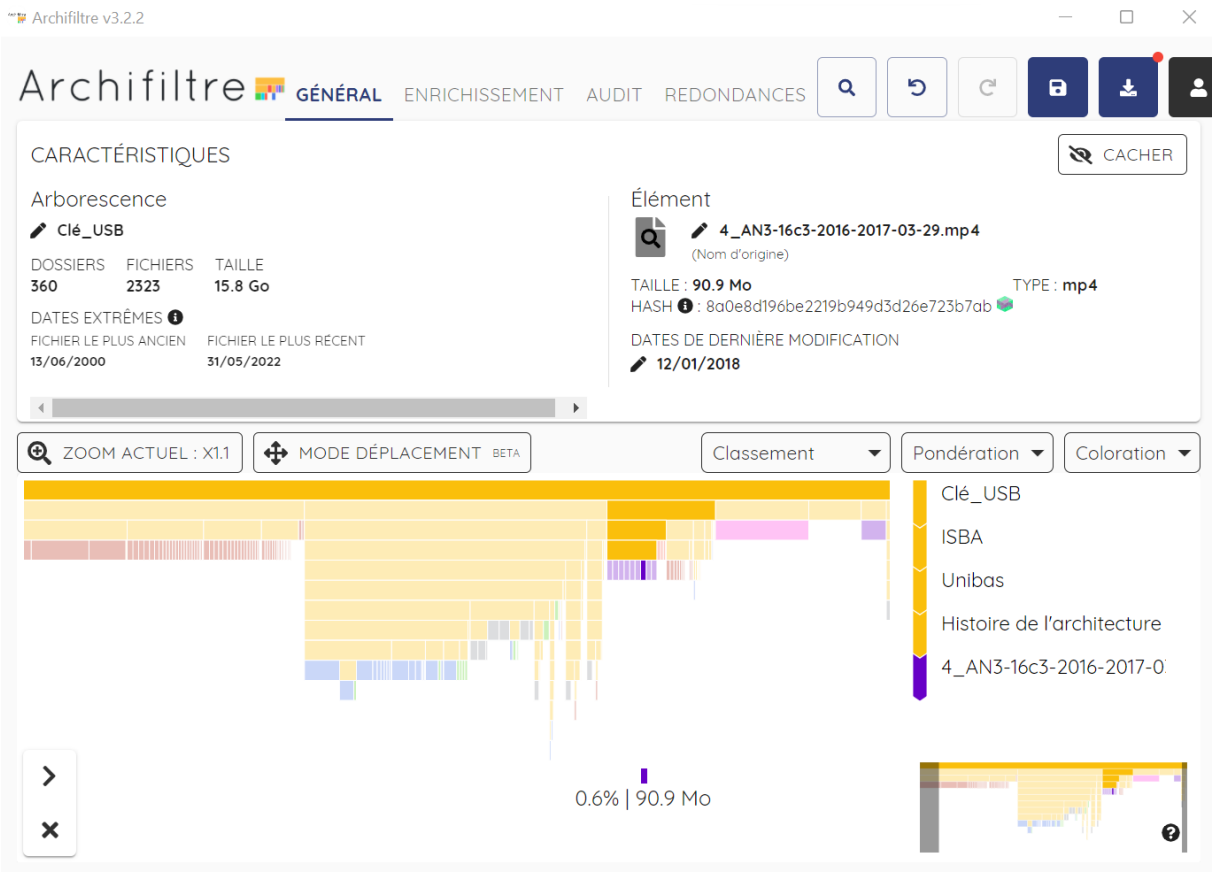
La fonctionnalité de visualisation de l'arborescence sera centrale dans l'étape de la « cartographie » (voir 5.2.1.4). Divers outils sur le marché proposent des fonctionnalités utiles pour l'appréhension d'une collection de fichiers peu structurée. S'il n'est pas dans les objectifs de ce travail de faire une présentation complète et ni de comparer entre eux ces logiciels[71], nous donnerons tout de même quelques exemples à titre d'illustration.

Ainsi le logiciel *Archifiltre*[72] propose une visualisation « en stalactite » (Figure 32), prenant le parti pris de ne pas représenter l'arborescence de gauche à droite, mais de haut en bas. Le dossier racine se trouve en haut et prend toute la largeur de la fenêtre, chaque niveau de profondeur des dossiers (en jaune) se développant par le bas, la largeur étant proportionnelle à la taille (en poids ou en nombre de fichiers contenus) du dossier, à l'extrémité des stalactites se trouvent les fichiers représentés par un code couleur par type de fichiers (bleu pour texte, rouge pour présentation, vert pour tableaux etc.).

[71] Pour cela, consulter l'article *Trois outils contribuant à l'archivage numérique* (2019) de Dominique Naud, l'étude de cas de Leonie Fritz (2021) réalisé dans le cadre du CAS ALIS ou le mémoire de recherche *Automatisation des fonctions archivistiques pour les données textuelles : quels outils et quelles fonctionnalités pour l'archiviste ?* (2022) de Bavaud et al. (HEG-GE).

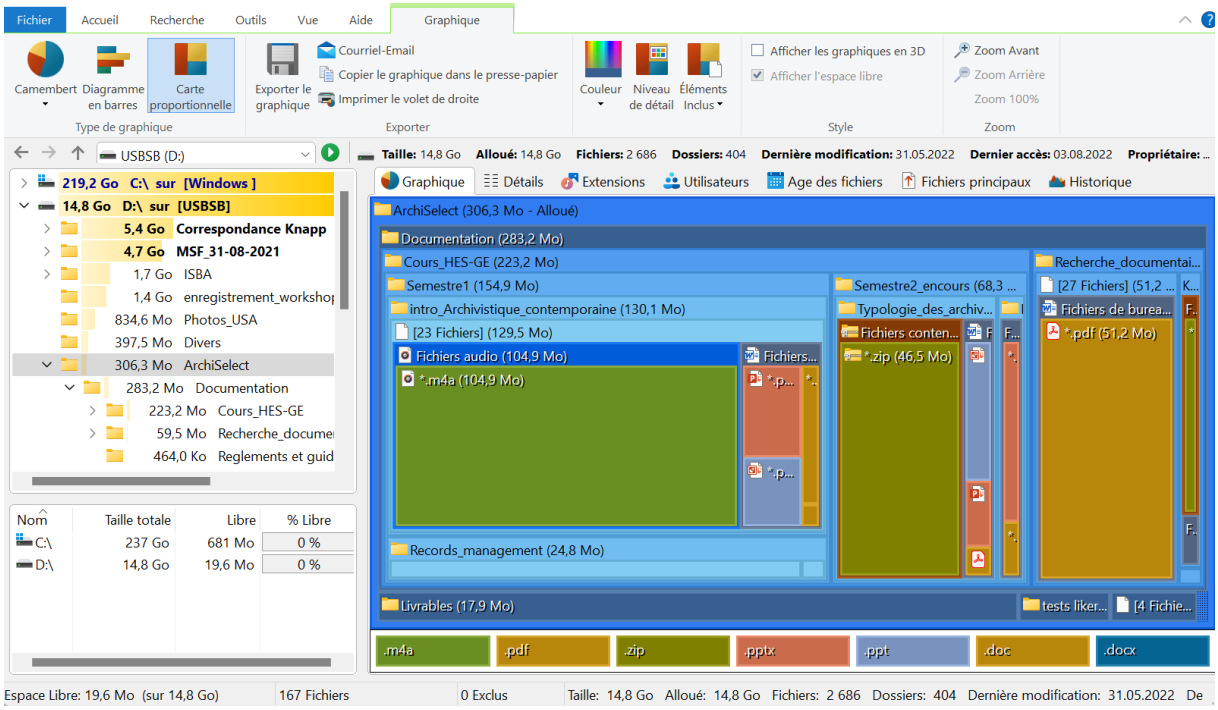
[72] https://archifiltre.fabrique.social.gouv.fr/ [Consulté le 3 août 2022]. La version 3.2.2 sur Windows a été utilisée pour créer la capture d'écran.

Figure 32 : Visualisation d'arborescence "en stalactite" dans *Archifiltre*



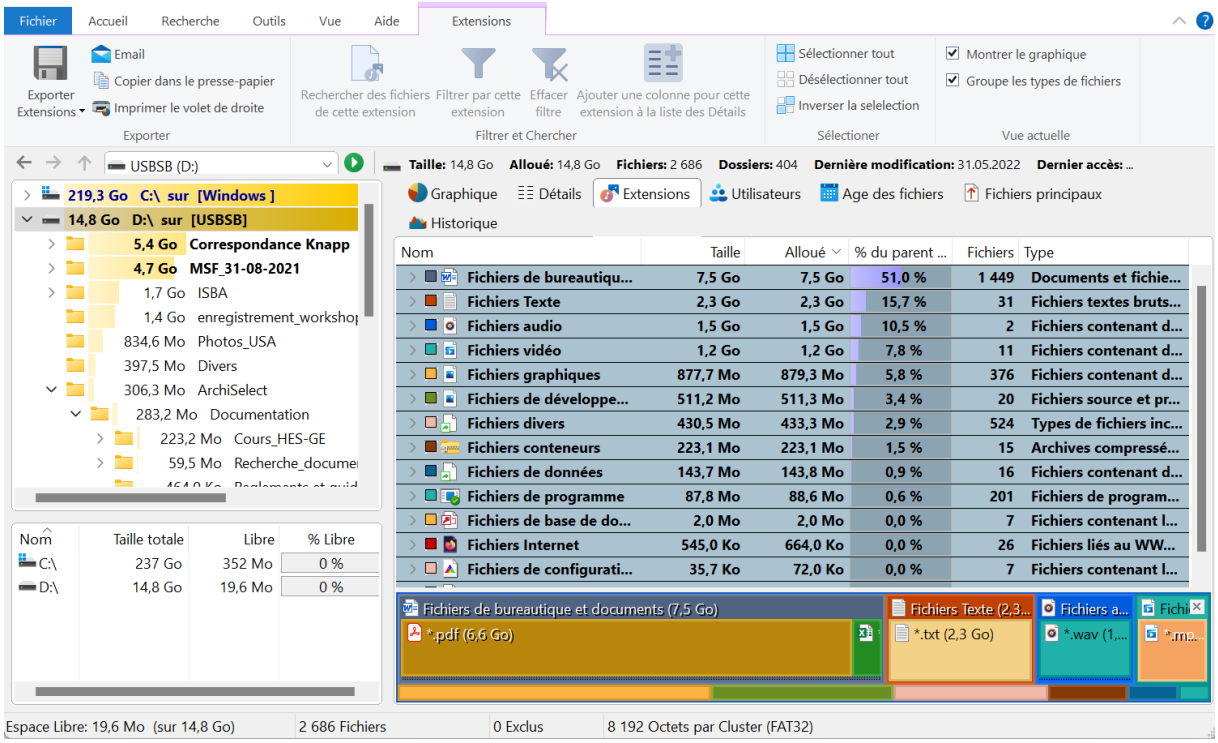
Cette représentation a le principal avantage de fournir rapidement une vision d'ensemble de la structure et de la volumétrie d'un jeu de données – qui sont les deux problématiques majeures dans l'appréhension d'une collection de fichiers non structurée.

Figure 33 : Visualisation d'arborescence en "carte proportionnelle" dans *TreeSize Pro*



TreeSize Pro[73] est un outil régulièrement cité dans les études de cas d'évaluation archivistique (Belovari 2017; Lenartz 2020; Shallcross 2015; Cocciolo 2016; Fritz 2021), s'il a été développé avant tout pour une gestion optimisée de l'espace disque, il offre une série de fonctionnalités utile pour l'archiviste appréhendant une collection de fichiers non structurée, à commencer par une représentation visuelle de « carte proportionnelle » (Figure 33), permettant une bonne visualisation de la volumétrie. Le logiciel permet également toute une série de recherches par extension, âge et taille des fichiers (Figure 34).

Figure 34 : Visualisation par format dans *TreeSize Pro*



Dans l'optique du développement d'un outil d'évaluation comme *ArchiSelect*, une réflexion plus poussée sur la visualisation pourrait être menée d'un point de vue de la structure. Par exemple, il serait intéressant de savoir plus clairement à quelle profondeur de l'arborescence se trouvent – en proportion – le plus de fichiers (soit la fin d'une arborescence) afin d'avoir une indication d'à quelle profondeur il est pertinent de débuter son travail d'évaluation.

Une autre piste serait de croiser la visualisation d'une arborescence avec d'autres indicateurs que les « classiques » volumétrie, format et (plus rarement) âge, en impliquant par exemple des métadonnées de contenus tels des entités nommées ou des thématiques (de type *topic modeling*). L'archiviste pourrait de cette manière se concentrer sur certains dossiers ou branches d'arborescence en laissant de côté d'autres qui sont hors-sujets.

Enfin, et pour faire le lien avec la prochaine partie, il pourrait être intéressant d'avoir la possibilité de visualiser les doublons (ou quasi-doublons) au sein d'une arborescence, cela permettrait de faire apparaître certains « motifs » laissés par les personnes ayant créé les dossiers – un dossier qui aurait été copié et dont les deux versions auraient continué à être alimentées par exemple.

[73] https://www.jam-software.com/treesize [Consulté le 15 août 2022]

5.3.2.2 Déduplication (f13)

L'importance de la déduplication n'est plus à démontrer, si les estimations varient fortement, entre 20 et 50% des documents détenus par les institutions et proposés aux archives seraient des doublons (Belovari 2017; The National Archives UK 2016). Même si nous nous basons sur les estimations les plus basses – et que l'on peut espérer que dans un environnement d'une maturité de gestion documentaire élevée ce pourcentage soit moindre – il reste que nous parlons là d'un volume gigantesque de données *a priori* sans valeur archivistique et facilement éliminables.

La question de la déduplication (ou dédoublonnage) mériterait sans doute un travail de recherche en soi. En effet, derrière une idée très simple se cache une réalité plus complexe. En premier lieu car même si deux fichiers sont techniquement parfaitement identiques, ce n'est pas pour autant qu'ils ne font pas « sens » aux deux lieux de dépôt, les développeurs d'*Archifiltre* préfèrent d'ailleurs parler de « redondances » plutôt que de doublons.

Dans la pratique, la prise de décision de quel « doublon » est à éliminer n'est ainsi ni anodine ni facile, et l'automatisation du dédoublonnage complexe. À titre d'exemple, des logiciels tel *Archifiltre* ou *TreeSize*, qui promettent pourtant des fonctionnalités de déduplication, demandent de sélectionner – de manière plus ou moins ergonomique – pour chaque doublon, lequel est à éliminer (Figure 35 et Figure 36), ce qui peut s'avérer extrêmement chronophage, même sur un petit jeu de données. Une piste à explorer serait ainsi une fonctionnalité de « création de règles » définissant sur un nombre de critères, quel doublon serait à éliminer.

Figure 35 : Fonctionnalité de déduplication dans *Archifiltre*

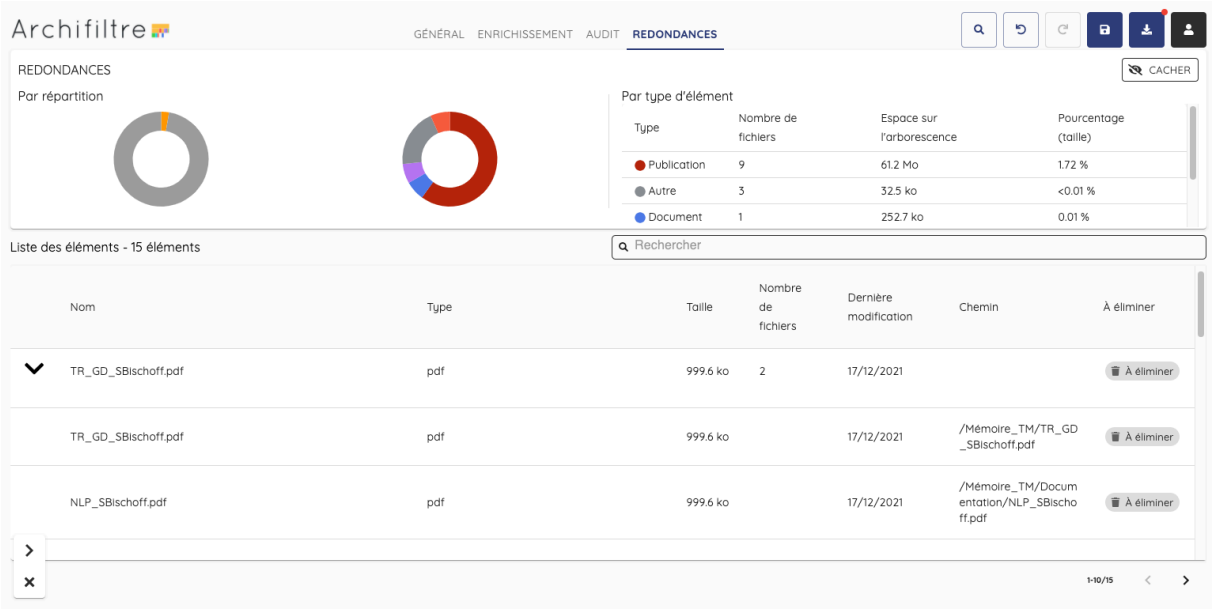
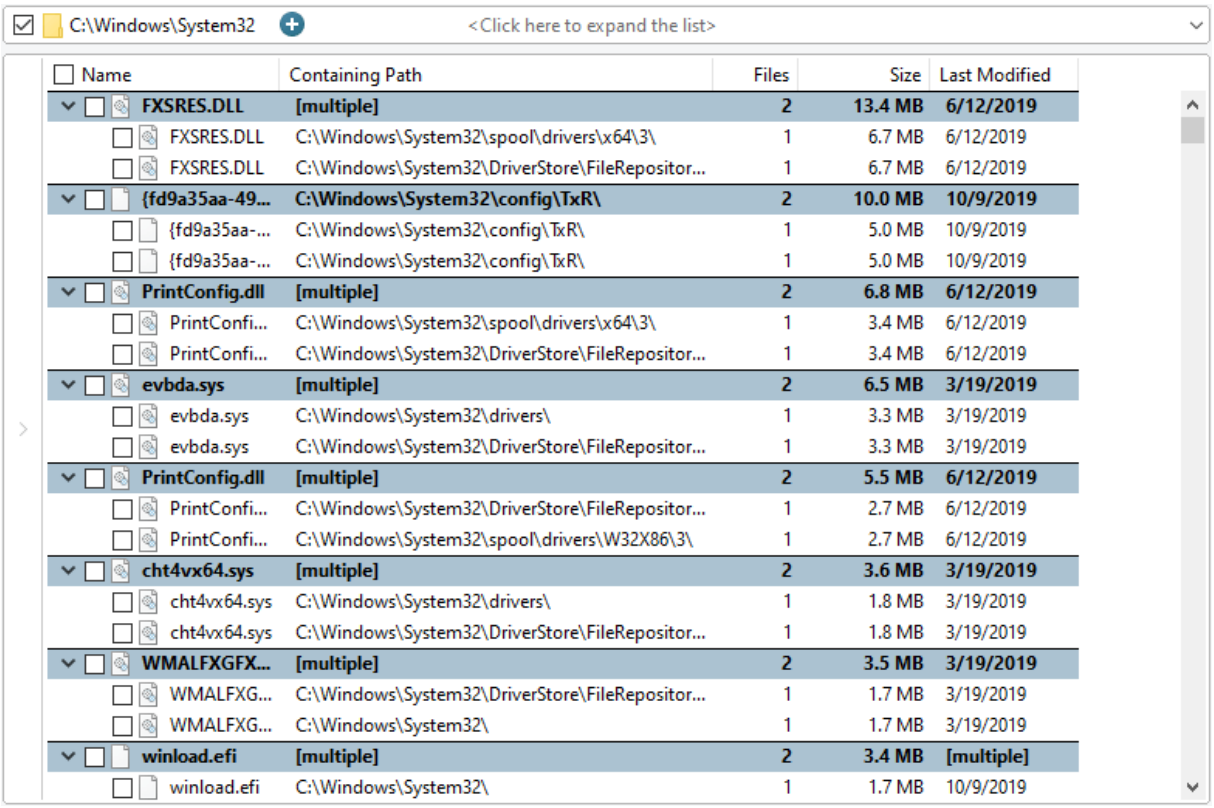


Figure 36 : Déduplication dans *TreeSize*



Name	Containing Path	Files	Size	Last Modified
FXSRES.DLL	**[multiple]**	**2**	**13.4 MB**	**6/12/2019**
FXSRES.DLL	C:\Windows\System32\spool\drivers\x64\3\	1	6.7 MB	6/12/2019
FXSRES.DLL	C:\Windows\System32\DriverStore\FileRepositor...	1	6.7 MB	6/12/2019
{fd9a35aa-49...	**C:\Windows\System32\config\TxR**	**2**	**10.0 MB**	**10/9/2019**
{fd9a35aa-...	C:\Windows\System32\config\TxR\	1	5.0 MB	10/9/2019
{fd9a35aa-...	C:\Windows\System32\config\TxR\	1	5.0 MB	10/9/2019
PrintConfig.dll	**[multiple]**	**2**	**6.8 MB**	**6/12/2019**
PrintConfi...	C:\Windows\System32\spool\drivers\x64\3\	1	3.4 MB	6/12/2019
PrintConfi...	C:\Windows\System32\DriverStore\FileRepositor...	1	3.4 MB	6/12/2019
evbda.sys	**[multiple]**	**2**	**6.5 MB**	**3/19/2019**
evbda.sys	C:\Windows\System32\drivers\	1	3.3 MB	3/19/2019
evbda.sys	C:\Windows\System32\DriverStore\FileRepositor...	1	3.3 MB	3/19/2019
PrintConfig.dll	**[multiple]**	**2**	**5.5 MB**	**6/12/2019**
PrintConfi...	C:\Windows\System32\DriverStore\FileRepositor...	1	2.7 MB	6/12/2019
PrintConfi...	C:\Windows\System32\spool\drivers\W32X86\3\	1	2.7 MB	6/12/2019
cht4vx64.sys	**[multiple]**	**2**	**3.6 MB**	**3/19/2019**
cht4vx64.sys	C:\Windows\System32\drivers\	1	1.8 MB	3/19/2019
cht4vx64.sys	C:\Windows\System32\DriverStore\FileRepositor...	1	1.8 MB	3/19/2019
WMALFXGFX...	**[multiple]**	**2**	**3.5 MB**	**3/19/2019**
WMALFXG...	C:\Windows\System32\DriverStore\FileRepositor...	1	1.7 MB	3/19/2019
WMALFXG...	C:\Windows\System32\	1	1.7 MB	3/19/2019
winload.efi	**[multiple]**	**2**	**3.4 MB**	**[multiple]**
winload.efi	C:\Windows\System32\	1	1.7 MB	10/9/2019

(Jam Software s.d.)

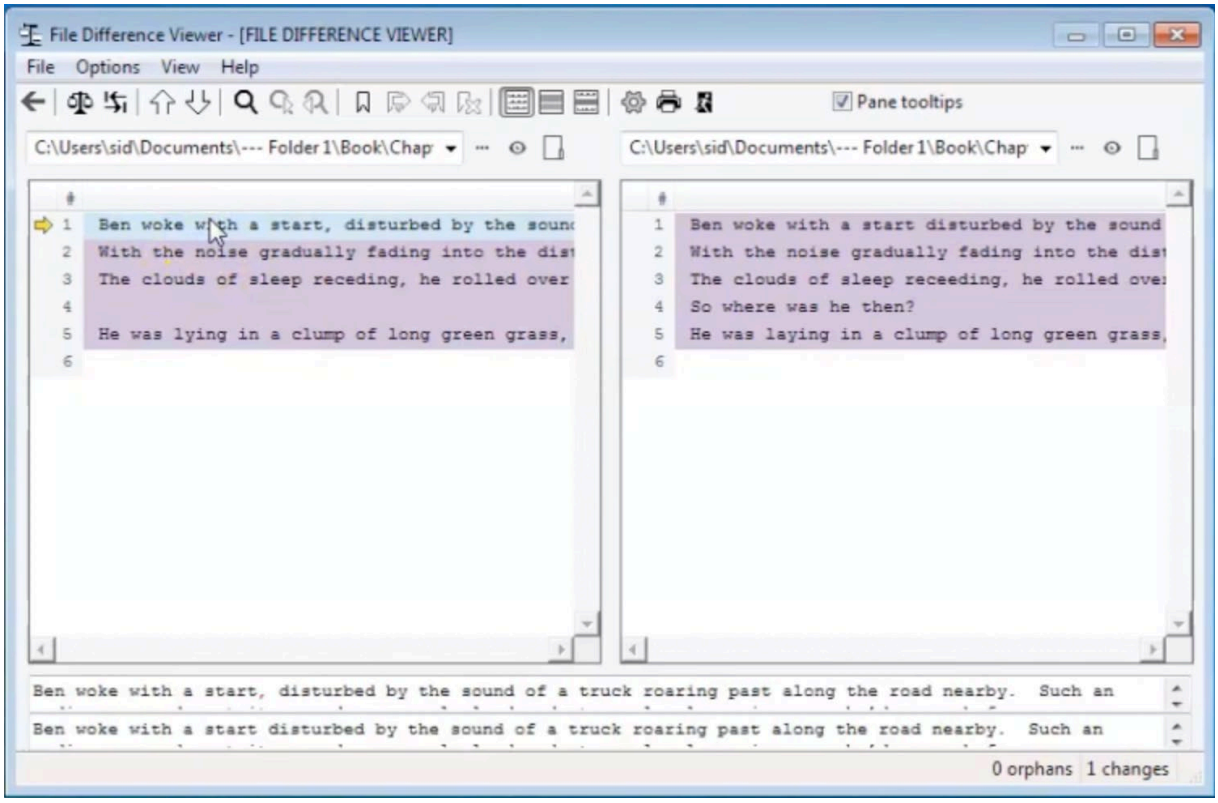
Une fois qu'il a été défini quel doublon va être éliminé, la problématique de la gestion du vide que celui-ci va laisser est réelle. Une piste est ce que *TreeSize* appelle les *hardlinks*, une de « fiche-fantôme » du fichier éliminé. Dans le cas concret du logiciel *TreeSize*, Susanne Belovari émet quelques réserves, tous les logiciels ne reconnaissant pas ces *hardlinks*, ce qui pourrait s'avérer problématique. L'auteure est par ailleurs de l'avis qu'à l'instar des pratiques d'évaluation de formats analogiques, peu de contexte significatif est perdu lors de désherbage et que la perte est pour le moins « acceptable » (2017, p. 59). Le choix d'éliminer les doublons doit, toujours selon l'auteure, prendre en compte la nature de la proposition – dans certains cas la déduplication n'est pas adaptée, elle est rejointe dans ce cas par Michael Shallcross (2015) qui avance une approche *MPLP*, la déduplication des fichiers spécifique peut être évitée et uniquement appliquée à des dossiers entiers de doublons.

En général sont utilisées des fonctions de hachage cryptographique pour la déduplication, qui permet d'attribuer à un document une valeur de hachage (*hash*) ou empreinte qui permet de démontrer la fixité d'un document, soit s'il a subi des modifications, les algorithmes les plus fréquents pour ce faire sont le MD5 et SHA1 (Spencer 2017). Ces valeurs d'empreintes sont également utilisées pour définir les documents identiques, soit des doublons. Dans le contexte strict de l'évaluation archivistique, il est cependant également intéressant de détecter les quasi-doublons[74], il en va d'ailleurs de même pour les dossier où il peut être intéressant de comparer les différences exactes, ce que proposent certains logiciels tel *Beyond compare*[75] (Figure 38) ou *FolderMatch*[76] (Figure 37). Cependant ces fonctionnalités nous semblent peu

[74] Pour plus de spécifications techniques, voir l'article de Ross Spencer (2017, p. 82-83).
[75] https://www.scootersoftware.com/index.php [Consulté le 3 août 2022]
[76] http://www.foldermatch.com/ [Consulté le 25 juin 2022]

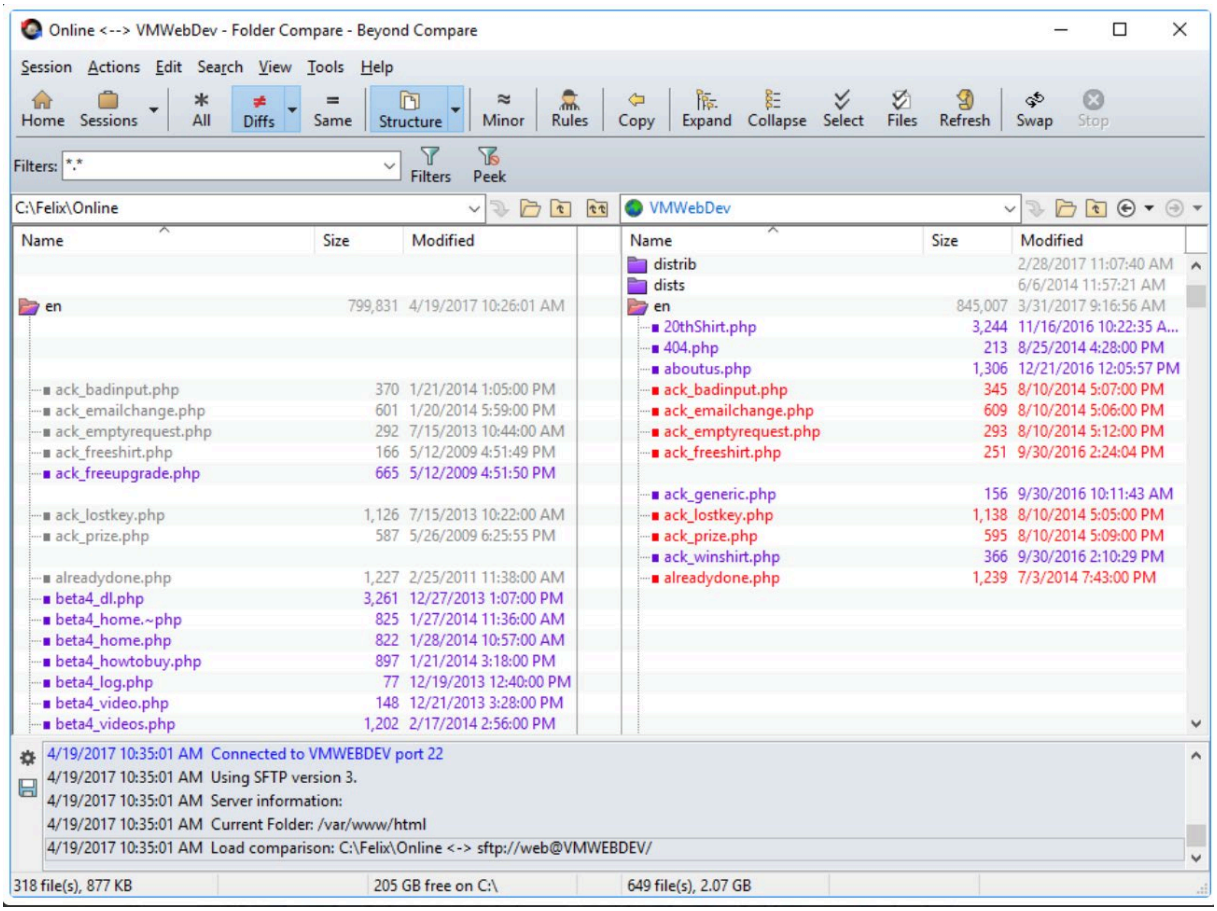
compatibles avec une évaluation archivistique telle que pratiquée aux AEN, si ce n'est pour une évaluation « approfondie », *in-depth appraisal* (Belovari 2017), d'un fonds privé par exemple.

Figure 37 : Comparaison de fichiers dans *FolderMatch*



(FolderMatch 2020)

Figure 38 : Comparaison de dossiers dans *Beyond Compare*



(Scooter Software 2022)

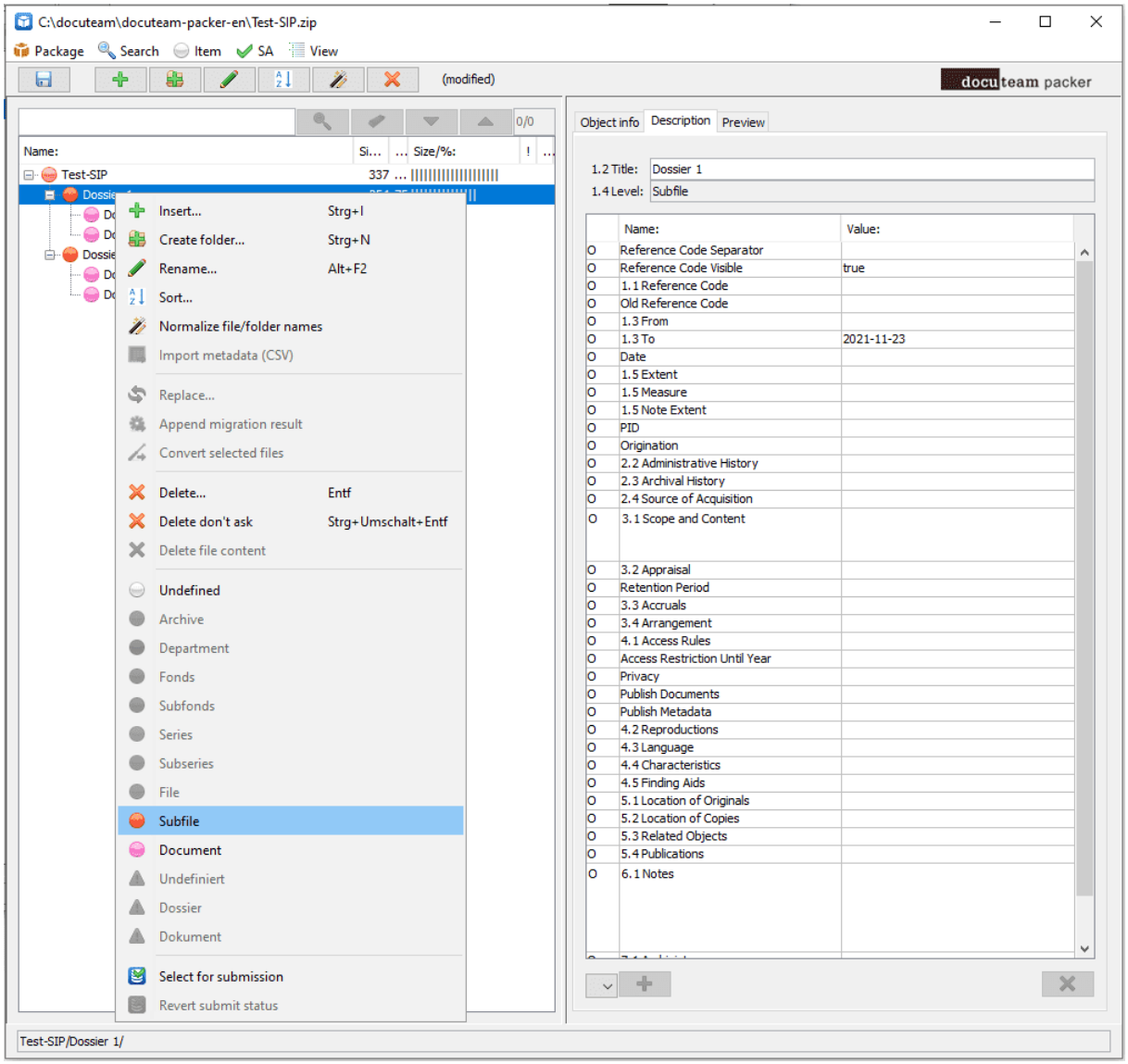
5.3.2.3 Réorganiser et renommer automatique (f31 et f32)

Il y a la volonté de la part des AEN[77] qu'*ArchiSelect* propose des fonctionnalités de réorganisation d'arborescence et de renommage automatique des fichiers. En effet, certains lots de données trop peu structurés vont devoir être retravaillés avant de pouvoir en tirer des SIP avec une granularité qui ait du sens pour les utilisateurs-trices. Divers outils actuels proposent des fonctionnalités semblables, tel *Archivematica* (voir 3.2.2.5), *Octave*[78] ou *docuteam packer*[79]. Tous ces outils sont pensés pour la création de SIP, offrant la possibilité d'ajouter des métadonnées descriptives (voir interface de droite, Figure 39), processus qui dans le cadre de la *SuiteArchi* ne serait pas prévu dans le périmètre d'*ArchiSelect* mais d'*ArchiPeren*.

La possibilité de sortir le reclassement et renommage du périmètre d'*ArchiSelect* et de l'inclure dans *ArchiPeren* serait, selon nous, plus en adéquation avec le concept AENeas, où chaque outil de la *SuiteArchi* couvre une étape du cycle de vie et doit pouvoir fonctionner indépendamment l'un de l'autre. De plus, le reclassement et renommage ne fait pas partie de la fonction archivistique de l'évaluation *stricto sensu*.

[77] Entretien avec le « Pôle évaluation » des AEN, le 11 juillet 2022, Neuchâtel.
[78] Outil de Constitution et de Traitement Automatisé des Versements Électroniques (OCTAVE) https://francearchives.fr/fr/article/88482499 [Consulté le 11 août 2022]
[79] https://docs.docuteam.ch/packer/6.3/en/index [Consulté le 11 août 2022]

Figure 39 : Réorganiser une collection de fichiers sur *docuteam packer*



(Docuteam 2022)

Une fonctionnalité proposée par le logiciel *WinCatalog*[80] est l'ajout de « dossiers virtuels », de manière qu'il n'y ait aucune inférence avec les données d'origine : l'archiviste crée un classement *virtuel* et y ajoute dossiers ou fichiers « *without losing a connection to the physical disk structure* » (WinCatalog 2022). Une telle fonctionnalité nous semble tout à fait en adéquation avec le concept *AENeas*, mais sa compatibilité avec le modèle OAIS n'est pas garantie. L'image disque (*disk image*) réalisée lors de l'étape de mise en sécurité des données (voir 5.2.1.2) serait cependant absolument à inclure dans le SIP.

5.3.2.4 Fonctionnalités impliquant de l'intelligence (AI)

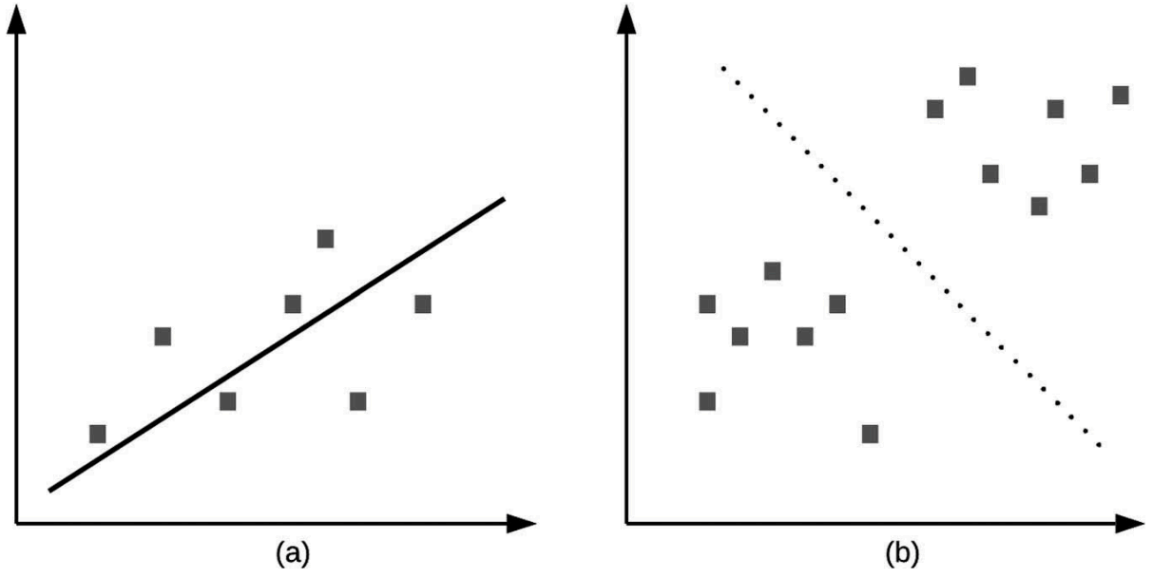
Nous l'avons évoqué en début de ce travail, sous intelligence artificielle (AI), nous comprenons avant tout des modèles statistiques, tel le *machine learning* (ML). Les fonctionnalités qui suivent impliquent du ML. En préambule, il importe de rappeler que ces modèles statistiques n'« interprètent » ni textes, ni images, les données doivent toujours être préparées et transformées, généralement en vecteurs afin de les rendre interprétables par la machine, une

[80] https://www.wincatalog.com/ [Consulté le 11 août 2022]

charge de travail non négligeable qu'il s'agira de prendre en compte dans le processus d'évaluation.

Si l'on se permet de schématiser, le ML tel qu'utilisé dans le domaine des archives cherche le meilleur moyen de diviser les données en différentes catégories ou *clusters* (Figure 40), ce sera à l'humain de décider du nombre de catégories et de quels *sens* ils ont : type de document, thème (*topic*) – voire sort final. Le grand défi du ML est d'identifier quelles caractéristiques influent directement sur le résultat. Ces caractéristiques ne se retrouvent en général pas dans les données brutes mais à un haut niveau d'abstraction, les outils de *deep learning* sont ainsi conçus pour classifier ces motifs dans les données brutes[81] (Rolan et al. 2019, p. 183).

Figure 40 : Ajustements de formules mathématiques à des points de données



(Rolan et al. 2019, p. 183)

Le *Natural Language Processing* (NLP) est un domaine multidisciplinaire à la croisée de la linguistique et du ML dont l'objectif est de faire comprendre le langage naturel (humain) à l'ordinateur. Pour ce faire, on prépare les données (texte) afin qu'elles soient lisibles par les machines, le *preprocessing*[82] une fois le texte transformé en vecteurs, les modèles statistiques du ML préalablement nourris de données d'entraînement peuvent être appliqués. Le NLP est à la base des fonctionnalités qui suivent.

5.3.2.4.1 Topic modeling (f23)

Le *Topic modeling* (littéralement la « modélisation de thèmes ») n'est autre qu'un modèle statistique qui découvre des redondances (thèmes) au sein d'un corpus de texte et les regroupe en *clusters*. Un grand intérêt du *topic modeling* est qu'il se fait en apprentissage non supervisé, il n'y a donc pas besoin de labelliser préalablement les documents. Le modèle le

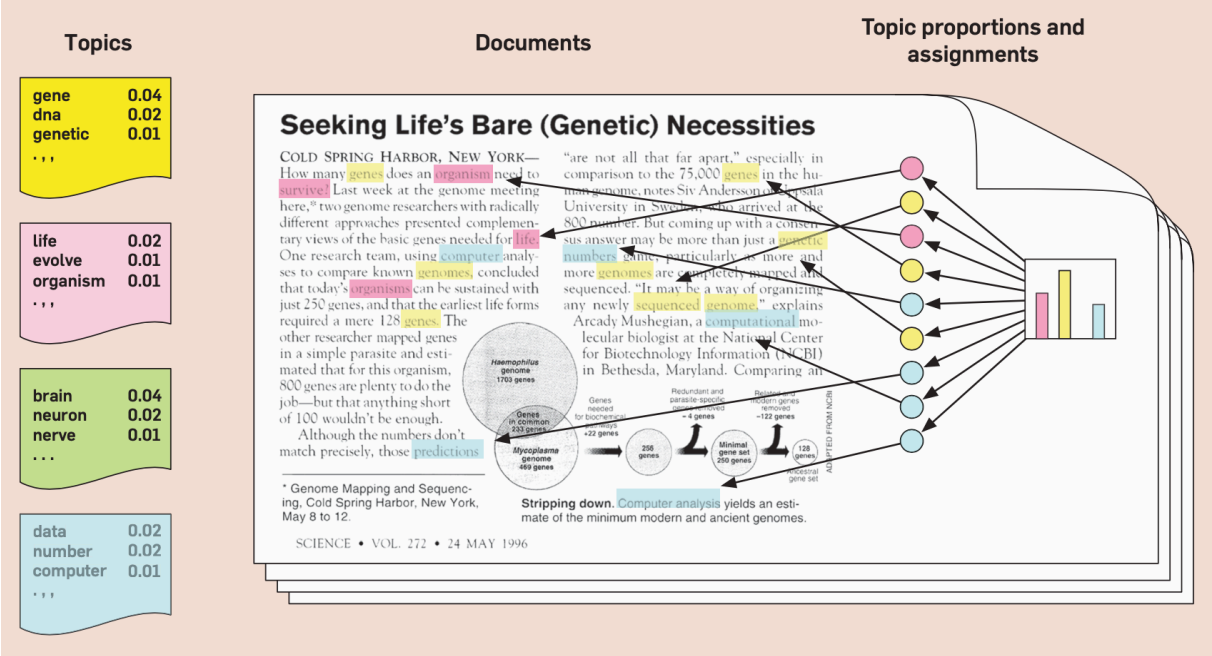
[81] Nous n'irons pas plus loin dans ce sujet complexe, pour une entrée en matière sur le *machine learning* et le *deep learning* voir Goodfellow et al. (2016).

[82] Cette étape passe en général par le la *tokenzation* qui segmente le texte en unités faciles à analyser ; le *stop word removal* qui retire les mots courts les plus fréquents ; le *lemmatization* et *stemming* qui réduit les mots à leur racine ; puis le *part-of-speech tagging*, qui attribue à chaque mot d'une phrase sa fonctionnalité grammaticale (Lutkevich, Burns 2021)

plus simple du *topic modeling* est la *Latent Dirichlet allocation* (LDA) (Blei, Ng, Jordan 2001) dont l'*intuition* est qu'un document comporte plusieurs thématiques, soit une distribution de mots dans le document (Blei 2012) (Figure 41).

De manière peu intuitive, le modèle n'indique pas un nombre de thèmes, c'est à l'archiviste d'en indiquer un nombre, et le modèle divisera le corpus en autant de *clusters* que demandé. Le modèle nous indiquera premièrement une liste de mots associés à un thème, du plus fréquent au moins fréquent et deuxièmement, l'attribution pour chaque document à un thème, un document pouvant faire partie de plusieurs thèmes à des pourcentages différents.

Figure 41 : *Topic Modeling* par *Latent Dirichlet Allocation*



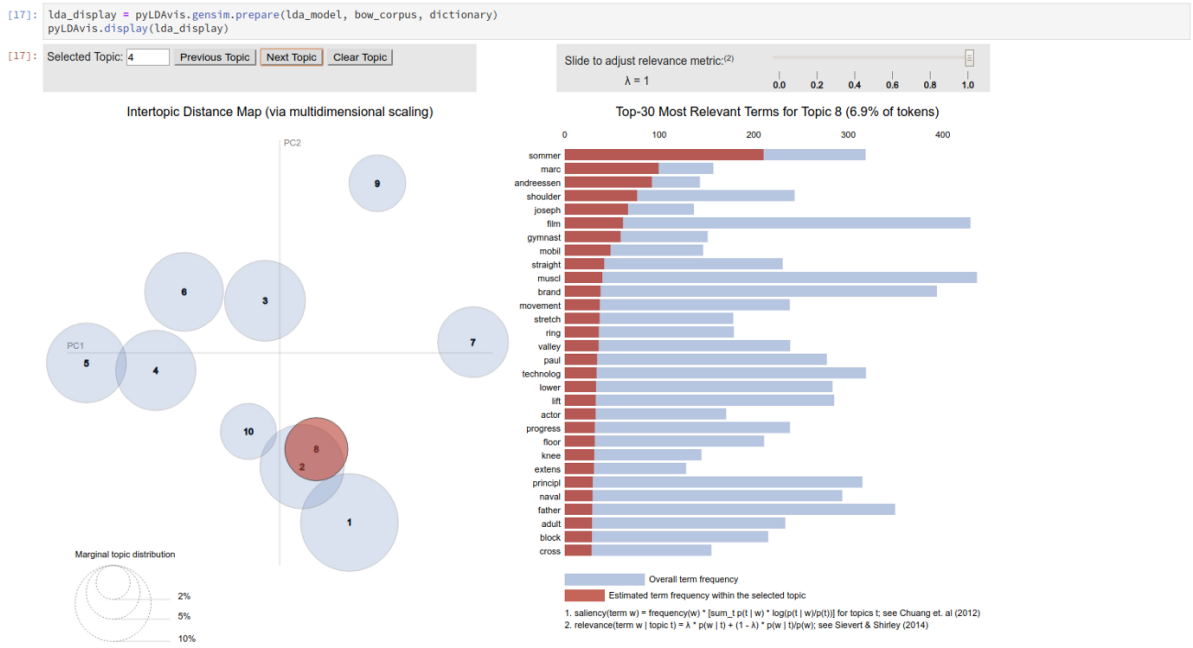
(Blei 2012, p. 78)

L'intérêt – du moins théorique – d'une telle technologie dans le cadre de l'évaluation archivistique semble manifeste. Le *topic modeling* offre un raccourci permettant à l'archiviste de prendre connaissance d'un grand nombre de documents sans les lire ni même les survoler.

Dans la dernière décennie, plusieurs projets ont exploité et testé l'application du NLP et notamment du *topic modeling*, quasi tous restant cependant au niveau du prototype. Morgan Goodman (2019) teste dans son travail la solution *open source BitCurator NLP* et son application de *topic modeling* : *bitcurator-nlp-gentm* et l'outil de visualisation *pyLDAvis*. Pour une utilisation efficace des modèles, une visualisation qui permet l'interaction avec l'utilisateur-trice est nécessaire.

L'outil de visualisation de *pyLDAvis* (Figure 42) permet ainsi de voir les thèmes et les relations entre les thèmes dans l'interface de gauche (plus ils sont éloignés, moins ils ont de rapport entre eux) et la prévalence (plus ils sont grands, plus le thème et présent dans le corpus). Une fois un thème sélectionné (ici le thème 8), s'affichent les barres rouges dans la partie de droite, indiquant la fréquence du terme au sein du thème (Goodman 2019; Sievert, Shirley 2014).

Figure 42 : Visualisation de *topic modeling* dans *pyLDAvis*



(Angelov 2018)

Si la technologie est prometteuse, la question de l'*interprétabilité* de ces thèmes est réelle. Dans l'étude de Goodman, les participants admettent qu'en situation réelle, sans savoir ce qui se trouve dans le jeu de données, il est difficile de se faire une idée en regardant uniquement les thèmes : « *I am not sure how this would help me analyze the collection if I didn't already know about the collection* » (2019, p. 28).

Autre problème de taille relevé par Goodman est, pour la solution testée[83], l'impossibilité d'appliquer l'analyse au seul document, l'analyse se faisant sur l'ensemble du corpus. D'autre part, l'extraction des fichiers s'avère extrêmement chronophage, une problématique commune aux technologies impliquant du ML, Goodman recommande ainsi son utilisation uniquement de manière ciblée (2019, p. 31).

Enfin, la validité même des *topic models* est sujette à critique, notamment dû aux variations de l'utilisation des mots à travers le temps (Schmidt 2012), mais son utilisation, couplée à la reconnaissance d'entités nommées offre à l'archiviste et au chercheur de grandes possibilités pour l'exploration et la description de matériaux numériques (Elings 2016).

5.3.2.4.2 Reconnaissance d'entités nommées (f21)

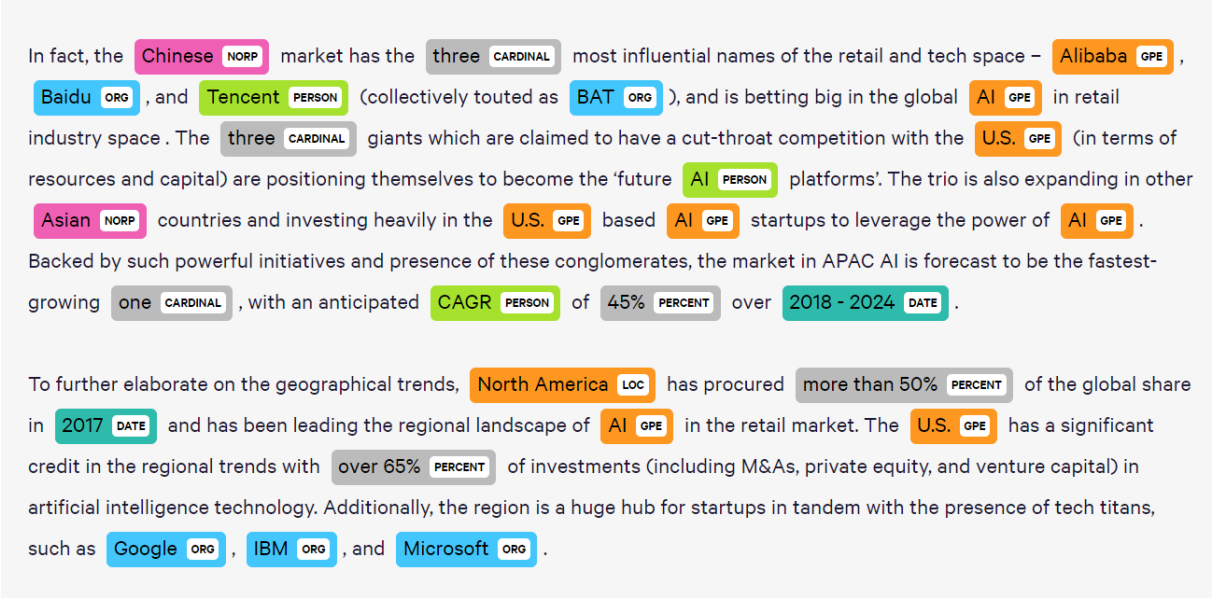
Avec le *topic modeling*, la *Named entity recognition* (NER), soit la reconnaissance d'entités nommées est la principale application du NLP dans le domaine de l'archivistique. Principalement appliquée dans le domaine de la recherche d'information (*information retrival*)

[83] Il existe cependant pléthore de solutions *open source* tel *Open NLP* de *Apache* https://opennlp.apache.org/ [Consulté le 19 juin 2022], le *package Topic models* de *R* https://cran.r-project.org/web/packages/topicmodels/index.html [Consulté le 19 juin 2022] ou *Mallet* https://mimno.github.io/Mallet/index [Consulté le 5 août 2022]. Ainsi que des solutions propriétaires : *Amazon Comprehend, IBM Watson, Google Cloud NLP, Aylien* etc. Concernant les solutions propriétaires de *machine learning* en général, se référer à l'étude de *The National Archives UK* (2020; 2021b; 2021a).

soit l'indexation de documents. En effet, diverses études indiquent que les pratiques de recherche d'information des historiens-nes impliquent des entités nommées (Duff, Johnson 2002; Duff, Harris 2002; Gooding 2016). D'autre part, la NER est largement utilisée en lien avec le *linked data* (données liées) permettant une indexation et une mise en relation de collections hétérogènes et multilingues à partir de fichiers d'autorités (Clough et al. 2011; Wilde, Hengchen 2016; Colavizza, Ehrmann, Bortoluzzi 2019; Ehrmann et al. 2021; Hawkins 2021). Dans le cas d'une potentielle implémentation d'une fonctionnalité NER dans *ArchiSelect*, les données ainsi générées pourraient par ailleurs profiter de fichiers d'autorités d'initiatives locales tel le projet *WikiNeocomensia*[84] visant à combler les lacunes d'articles, données et informations du patrimoine neuchâtelois dans *Wikipédia*.

L'utilisation de la NER spécifiquement pour l'évaluation archivistique est ainsi rarement abordée dans la littérature professionnelle, cependant le potentiel de la technologie nous semble bien réel, car elle offre des bras de leviers similaires au *topic modeling*, soit accéder rapidement à des informations de description de contexte d'un corpus en s'épargnant la lecture. D'autre part, le mandat réalisé par la HEG-GE prévoit également de mettre à contribution les informations ainsi récoltées pour accorder plus de confiance aux métadonnées.

Figure 43 : Reconnaissance d'entités nommées avec *SpaCy*



(Sharma 2021)

D'un point de vue purement technique, diverses solutions *open source* existent, tel *SpaCy*[85] (Figure 43), *OpenNLP*[86] ou *Stanford NER*[87], pour ne citer qu'eux. Pour un état des lieux complet de l'utilisation de la NER sur les archives historiques, voir Ehrmann et al. (2021).

5.3.2.4.3 Sensitivity review (f28)

La *sensitivity review*, est l'identification automatique de données sensibles. Divers logiciels ou applications impliquant du NLP propriétaires (The National Archives UK 2021a) et *open source*

[84] https://fr.wikipedia.org/wiki/Projet:WikiNeocomensia [Consulté le 11 août 2022]).
[85] https://spacy.io/ [Consulté le 11 août 2022]
[86] https://opennlp.apache.org/ [Consulté le 11 août 2022]
[87] https://nlp.stanford.edu/software/CRF-NER.shtml [Consulté le 11 août 2022]

(Hutchinson 2020, p. 159-161) promettent des fonctionnalités capables d'identifier au sein d'un corpus les données sensibles. En effet, si le besoin pour une telle technologie est réel, particulièrement dans le domaine publique (Baron, Payne 2017; McDonald, Macdonald, Ounis 2017; Gollins et al. 2014; Moss, Gollins 2017; Sloyan 2016), certains auteur-e-s, et nous les rejoignons, sont de l'avis que la technologie n'est pas encore mûre pour répondre à ce besoin. La définition même d'une donnée sensible est sujette à débat et absolument dépendante du contexte (Gollins et al. 2014; Moss, Gollins 2017). Nous recommandons ainsi l'implémentation d'une fonctionnalité d'identification des données sensibles en théorie, mais déconseillons, en l'état, l'utilisation de fonctionnalités de *sensitivity review* sur la base de NLP.

5.3.2.4.4 Classification automatique des documents (f22)

Nous n'entrerons pas ici dans une explication technique de la classification automatique de documents qui dépassent nos connaissances, mais souhaitons proposer une réflexion globale sur l'utilisation de la classification automatique. Le principe général de la classification automatique reste semblable à celui du *topic modeling* : une analyse statistique des éléments composant un document (au choix : du texte, des métadonnées de tout genre ou encore la mise en page) qui sont ensuite répartis à l'aide du *machine learning* en *clusters* de types de documents. Bien que contrairement au *topic modeling* il s'agirait d'un entraînement *supervisé*, se basant sur un jeu de données labellisé par un archiviste au préalable.

On a pu observer une certaine réticence du milieu de l'archivistique à faire recours à une automatisation complète (Esposito et al. 2004; Alberts, Vellino 2013; Vellino, Alberts 2016; The National Archives UK 2021a), soit l'équivalent de l'*autocategorization* dernier degré de l'automatisation énoncé par la typologie de la NARA (2014, p. 14). Bien quelques recherches ont été menées (Esposito et al. 2004; Alberts, Vellino 2013; Vellino, Alberts 2016; The National Archives UK 2021a), mais peu d'exemples de réussite clairs sont cependant à signaler (Rolan et al. 2019, p. 185).

De notre point de vue, plusieurs éléments vont contre l'utilisation du *machine learning* pour la classification automatique des documents, du moins dans le contexte du développement d'*ArchiSelect*. En effet, Rolan et al. démontrent bien qu'à l'heure actuelle, la complexité d'automatiser la compréhension du contexte (Parapadakis 2013) est encore trop importante pour pouvoir la déléguer aux modèles prédictifs, même aux plus puissants qui soient à disposition. Le problème principal restant la « transférabilité »[88] des modèles. En effet, des modèles potentiellement capables de subvenir aux besoins techniques et éthiques de l'archivistique « *may require complex, interdependent, multi-stage, deep learning algorithms that remain still within the realm of fundamental AI research* » (2019, p. 187).

Tant que l'archivistique n'aura pas une compréhension approfondie des technologies du ML, son implémentation nous semble problématique, l'appel de Marciano et al. de créer une « transdiscipline » de *computationnal archival science* serait une piste (2018) à suivre. Cependant, même si l'archiviste devient expert-e de science informatique, reste la question fondamentale de la *black box* que représente le *deep learning*. En effet, la technologie est

[88] Les auteurs invoquent le *no free lunch theorem* (Wolpert, Macready 1997), qui démontre qu'aucun modèle de prévision n'est mieux qu'un autre dès lors qu'il est appliqué à tous les jeux donnés possibles, même s'ils sont apparemment similaires « *any two optimization algorithms are equivalent when their performance is averaged across all possible problems* » (Wolpert, Macready 2005, p. 721)

capable de nous fournir des réponses – dans certains domaines bien plus précises que l'être humain – mais n'est, par principe, pas capable de nous dire *comment* il est arrivé à cette réponse. Bien que la recherche s'intéresse actuellement à l'interprétabilité des résultats des modèles prédictifs de type *deep learning* (Sun et al. 2021), il reste problématique, dans un souci de traçabilité des décisions, d'impliquer une technologie telle que le ML pour une fonction aussi centrale qu'est l'attribution d'un sort final à un document.

6. Discussion

Le changement de paradigme dans la volumétrie de la production documentaire ainsi que le manque de structure inhérent aux collections de fichiers sont deux défis majeurs de l'évaluation de documents numériques. Ce défi va être relevé dans le canton de Neuchâtel par le développement d'un outil d'aide à l'évaluation archivistique : *ArchiSelect*.

Dans ce travail, nous avons essayé d'explorer quelles fonctionnalités pourraient être implémentés dans le logiciel *ArchiSelect*, prenant en compte les spécificités du contexte neuchâtelois : le cadre légal, la situation de gestion documentaire au sein de l'administration ainsi que les pratiques d'évaluation en cours. En effet :

> *« Toute stratégie, quelle que soit sa performance, a besoin de s'adapter à son contexte pour mieux répondre à ses besoins. Les stratégies d'évaluation des archives en sont un exemple. Pour réussir leur mission, elles ont besoin de considérer les enjeux importants qui caractérisent leur contexte d'application. » (Makhlouf Shabou 2015b, p. 199)*

6.1 Résultats

L'analyse de la revue de littérature et l'état des lieux des projets et outils existants est sans appel : il n'existe actuellement aucune solution pratique et théorique faisant l'unanimité au sein de la profession. Les rares initiatives ayant abouti à des résultats prometteurs étaient basés sur des projets et sont restés sous forme de prototypes.

Le besoin d'automatisation est cependant manifeste. Cette automatisation devra selon toute vraisemblance passer par l'utilisation du *machine learning*, une forme d'intelligence artificielle. Cependant, bien que cette dernière se démocratise, une automatisation totale ne semble pas encore envisageable – voire souhaitée. Le domaine de l'archivistique devra investir le domaine des *computational sciences* et faire sien ses principes avant de pouvoir faire recours aux modèles prédictifs, pour des raisons de traçabilité et de reproductibilité.

Le cadre conceptuel proposé par Basma Makhlouf Shabou (2010; 2012a; 2012b; 2015a; Makhlouf Shabou et al. 2020) couplé au *text mining* tel que proposé dans le mandat de recherche mené par les Prof. Makhlouf Shabou et Gaudinat de la HEG-GE, offre des perspectives prometteuses d'automatisation qui ne couvre cependant pas toutes les dimensions d'évaluation. Notamment la *représentativité*, qui comprend les sous-dimensions centrales de *contexte organisationnel* et *socio-culturel* n'est pour ainsi dire pas automatisable.

D'autres solutions sous forme de fonctionnalités implémentables doivent ainsi être envisagées pour rendre possible l'évaluation de grands volumes peu structurés. Notre état des lieux a démontré que de telles solutions, bien qu'encore peu standardisés et avec une courbe d'apprentissage élevée, existent. En effet, les fonctionnalités tel le *topic modeling*, de reconnaissances d'entités nommées voire la classification automatique des documents peuvent être assemblés de manière cohérente en *wokflow*, nous pensons l'avoir démontré.

6.2 Limites de la recherche

Notre démarche de vouloir « faire émerger » les besoins de fonctionnalités de soutien à l'évaluation par la « traduction dans le numérique » des tâches effectuées dans la pratique d'évaluation de documents analogiques a tout de même montré ses limites. En effet, si les principes théoriques restent les mêmes, les principaux défis du numérique viennent notamment des différences avec les documents analogiques, d'autre part la présence (même

si peu systématique) de métadonnées et la lisibilité des documents par la machine offrent d'autres opportunités que les documents analogiques. Ainsi, une réflexion approfondie sur l'utilisation des métadonnées – notamment dans le contexte du *mapping* entre métriques archivistiques et métriques de fouille de données – aurait pu être mené dans ce travail.

D'autre part, si notre proposition de *workflow* pour l'évaluation de documents numériques est certes inspirée des pratiques d'évaluation ayant court à Neuchâtel, celui-ci reste *fictif* et va devoir être éprouvé et testé en situation réelle.

6.3 Recommandations

Notre recherche reste cependant valide car elle offre une approche certes théorique, mais basée sur une littérature récente et fournit ainsi une base de travail solide pour l'évaluation de documents numériques. En effet, les différentes fonctionnalités thématisées sont à l'heure actuelle proposées dans une large majorité par des outils propriétaires et *open source*. Et c'est notre principale recommandation : tester « à blanc » les différentes fonctionnalités assemblées en *workflow* sur des jeux de données réels issus des entités organisationnelles soumises à la LArch. Puis par itération adapter le *workflow* et les fonctionnalités utiles.

Un tel test à l'interne permettra également aux archivistes de s'approprier les technologies en question, ce qui facilitera à son tour grandement la rédaction du cahier des charges. L'aspect de l'interface graphique va également grandement être influencé par les choix d'implémentation de fonctionnalités. Enfin, cette étape permettra au demeurant de mettre en place une approche intégrée et les processus nécessaires à la réception des premiers versements numériques dans l'attente de la finalisation d'*ArchiSelect*.

D'autre part, une réflexion approfondie et qualitative doit d'après nous encore être menée sur les variables, sur leur sélection mais avant tout sur les « règles » proposées où il existe actuellement qu'une base de travail. L'utilisation du *machine learning*, domaine extrêmement volatile et en pleine expansion demandera lui aussi un état de l'art actualisé, réalisé par un spécialiste.

Enfin, dans le scenario d'une utilisation du ML et du NLP, nous ne pouvons qu'appeler à l'application des principes de conception (*design principles and workflow considerations*) suggérés par Hutchinson (2020, p. 166-168). Si certains paraissent évidents, tel la facilité d'utilisation, flexibilité ou *configurabilité*, les deux derniers nous semblent moins flagrants. D'une part l'*iterativity*, fortement liée à la facilité d'utilisation, soit la possibilité de *raffiner* de manière itérative les modèles prédictifs par une interface utilisateur. D'autre part, l'interopérabilité qui, d'après Hutchinson est une exigence fondamentale dans l'utilisation de NLP. En effet il plaide si ce n'est pour une approche d'outils spécialisés assemblés en *workflow*, pour des approches intermédiaires du type *BitCurator* ou *Archivematica,* compatibles avec des outils intégrés. L'interopérabilité est également importante afin de pouvoir partager des ressources tels des dictionnaires d'entités ou modèles prédictifs entre institutions semblables où la transférabilité est plus aisée.

7. Conclusion

Quelles fonctionnalités seront utiles à l'archiviste dans sa tâche d'évaluation de documents numériques ? Pour répondre à notre question de recherche, nous avons pris en compte d'une part le contexte documentaire neuchâtelois – marqué par une gestion documentaire hétérogène, une absence d'un réel *Records management* et un cadre légal dont le périmètre de l'obligation de proposer est assez large – et d'autre part les pratiques d'évaluation des Archives de l'État de Neuchâtel, marquées par une tradition latine très attachée au point de vue de la mémoire collective (Chabin, Watel 2006) et évitant un interventionnisme trop marqué auprès des entités productrices d'information (Cadre stratégique, voir 5.1.2). Partant de là, nous avons identifié et documenté un set de fonctionnalités-clés combinées au sein d'un *workflow* pour l'évaluation de documents numérique compatible avec les pratiques d'évaluation archivistique neuchâteloises.

Le mandat réalisé par la HEG-GE a fourni des solutions convaincantes pour l'évaluation des dimensions de *valeur de preuve* et d'*exploitabilité*. Cependant, notre analyse a démontré qu'*ArchiSelect* allait devoir inclure des fonctionnalités capables, si ce n'est de mesurer, du moins de fournir des indicateurs sur la dimension de *représentativité* des documents, dimension problématique par son aspect subjectif. Le *workflow* et le set de fonctionnalités proposés donnent quelques pistes, non d'automatisation pure, mais des « bras de leviers » pour l'évaluation de documents numériques. Le set comprend tant des fonctionnalités que l'on pourrait qualifier de classiques, tels la visualisation d'arborescences, l'identification des formats ou la déduplication, que des fonctionnalités plus expérimentales comme le *topic modeling* qui se base sur des modèles prédictifs de *machine learning*.

Aucune des fonctionnalités proposées ne sera cependant une solution miracle. En effet, le *machine learning*, dont l'appropriation de la part de la profession d'archiviste est certes en cours, mais doit encore être consolidée, a suscité énormément d'espoirs qui peinent encore à se concrétiser. L'AI n'est et ne sera sans doute jamais une « *silver bullet* » prête à l'emploi qui viendrait ôter les difficultés de l'évaluation archivistique (Rolan et al. 2019). Et dans la perspective de l'utilisation de l'AI pour la fonction d'évaluation, l'archiviste devra, plus que jamais, rester conscient de sa position de toute-puissance[89] et de la responsabilité sociale qui en découle. Faut-il pour autant considérer l'évaluation archivistique comme l'éternelle *faiblesse des tout-puissants*[90] ? Nous ne pensons pas. Les archives sont « infiniment humaines » (Yoakim 2022), et la pratique d'évaluation ne peut ainsi être dotée d'une dimension totalement objective voire scientifique (Ducharme 2000, p. 21), cependant elle réclame transparence et pondération (Coutaz 2016b, p.33), des principes qui vont devoir guider les choix de développement d'*ArchiSelect*.

[89] Tom Nesmith évoque « *the tension within the central archival professional myth: enormous power and discretion over societal memory, deeply masked behind a public image of denial and self-effacement* » (2002, p. 32)

[90] Formule empruntée à Alain Bashung, *Comme un Lego* (Bleu Pétrole, 2008)

Bibliographie

Ablage. *Leo*. [en ligne]. 2022. [Consulté le 9 juillet 2022] Disponible à l'adresse : https://dict.leo.org/allemand-fran%C3%A7ais/Ablage

AIMS WORK GROUP, 2012. *AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship* [en ligne]. Janvier 2012. University of Hull; Standford University; University of Virginia; Yale University. [Consulté le 13 juin 2022]. Disponible à l'adresse : https://dcs.library.virginia.edu/files/2013/02/AIMS_final_A4.pdf

ALBERTS, Inge et FOREST, Dominic, 2012. Email pragmatics and automatic classification: A study in the organizational context. *Journal of the American Society for Information Science and Technology* [en ligne]. 2012. Vol. 63, n° 5, pp. 904-922. [Consulté le 18 avril 2021]. Disponible à l'adresse : https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21702

ALBERTS, Inge et VELLINO, André, 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]. Disponible à l'adresse : https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/meet.14505001112

ANGELOV, Boyan, 2018. What do successful people talk about? A machine learning analysis of the Tim Ferris Show. *towards data science* [en ligne]. 15 décembre 2018. [Consulté le 25 novembre 2021]. Disponible à l'adresse : https://towardsdatascience.com/what-do-successful-people-talk-about-a-machine-learning-analysis-of-the-tim-ferris-show-161fc7ed4394

Apprentissage automatique. *Grand dictionnaire terminologique* [en ligne]. 2012. [Consulté le 25 juillet 2022]. Disponible à l'adresse : https://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=8395061

Apprentissage profond. *Grand dictionnaire terminologique* [en ligne]. 2012. [Consulté le 10 août 2022]. Disponible à l'adresse : https://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=26532876

ARCHIVES DE L'ÉTAT DE NEUCHÂTEL, [s.d.] a. SuiteArchi. *République et canton de Neuchâtel* [en ligne]. [s.d.]. [Consulté le 30 mai 2022 a]. Disponible à l'adresse : https://www.ne.ch/autorites/DESC/SCNE/archives-etat/numerique/Pages/76-BoiteOutils.aspx

ARCHIVES DE L'ÉTAT DE NEUCHÂTEL, [s.d.] b. Archivage numérique. *République et canton de Neuchâtel* [en ligne]. [s.d.]. [Consulté le 14 juillet 2022 b]. Disponible à l'adresse : https://www.ne.ch/autorites/DESC/SCNE/archives-etat/numerique/Pages/76-BoiteOutils.aspx

ARCHIVES DE L'ÉTAT DE NEUCHÂTEL, 2016. *Projet AENeas* : Description du projet AENeas basé sur le concept. Neuchâtel : Archives de l'État, novembre 2016

ARCHIVES DE L'ÉTAT DE NEUCHÂTEL, 2019. *Système de contrôle interne 325 Proposition et évaluation définitive* [fichier Word]. 2019. Document interne à l'institution

ARCHIVES DE L'ÉTAT DU VALAIS, 2020. *Politique d'acquisition des Archives de l'Etat du Valais (2021-2025)* [en ligne]. 22 avril 2020. [Consulté le 12 août 2022]. Disponible à l'adresse : https://www.vs.ch/documents/249470/7491968/Politique+d%27acquisition+AEV+2021-2025.pdf/662f52fa-9a5a-9001-1981-43a48ccf597f?t=1600082341278

ARCHIVES FÉDÉRALES SUISSES, 2022. Gestion des affaires. *Archives fédérales suisses* [en ligne]. 29 juin 2022. [Consulté le 9 juillet 2022]. Disponible à l'adresse : https://www.bar.admin.ch/bar/fr/home/informationsmanagement/geschaeftsverwaltung.html

ASSOCIATION DES ARCHIVISTES FRANÇAIS (éd.), 2012. *Abrégé d'archivistique: principes et pratiques du métier d'archiviste*. 3e éd. revue et augmentée. Paris : Association des archivistes français. ISBN 978-2-900175-03-3. 020

ASSOCIATION DES ARCHIVISTES FRANÇAIS, 2018. *Fiche pratique AMAE n°22 : Réflexion sur le vrac numérique* [en ligne]. [Consulté le 9 juin 2022]. Disponible à l'adresse : https://www.archivistes.org/IMG/pdf/fp22_reflexions_vrac_numerique.pdf?7512/9f51d7000239d6b41b3be59a7b0f2c3748adfeae

Automatisation. *Encyclopædia Universalis* [en ligne] 2022. [Consulté le 1er août 2022]. Disponible à l'adresse : https://www.universalis.fr/encyclopedie/automatisation/

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

BARON, Jason R. et PAYNE, Nathaniel, 2017. Dark Archives and Edemocracy: Strategies for Overcoming Access Barriers to the Public Record Archives of the Future. In : *2017 Conference for E-Democracy and Open Government (CeDEM)*. Mai 2017. pp. 3-11.

BARTOLINI, Lionel, 2021. Le faux document, une source historique de première importance. *arbido* [en ligne]. 2021. N° 4. [Consulté le 1 juin 2022]. Disponible à l'adresse : https://arbido.ch/fr/edition-article/2021/das-dokument/le-faux-document-une-source-historique-de-premiere-importance

BAUER, Philipp G., 1946. The Appraisal of Current and Recent Records. *National Archives Staff Information Circulars*. 1946. n°23.

BAVAUD, Aurélie, BUSSARD, Denis et BISCHOFF, Sébastien, 2022. *Automatisation des fonctions archivistiques pour les données textuelles : quels outils et quelles fonctionnalités pour l'archiviste ?* Mémoire de recherche. Genève : Haute École de Gestion Genève.

BEARMAN, David, 1991. An indefensible bastion: archives as repositories in the electronic age. In : BEARMAN, David, *Archival management of electronic records*. Pittsburgh: Archives & Museum Informatics.

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

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]. Disponible à l'adresse : https://doi.org/10.1080/15332748.2018.1503014

BISCHOFF, Frank M., 2014. Bewertung elektronischer Unterlagen und die Auswirkungen archivarischer Eingriffe auf die Typologie zukünftiger Quellen. *Archivar*. janvier 2014. Vol. 67, pp. 40-52.

BITCURATOR, [s.d.]. BitCurator Project. *BitCurator* [en ligne]. [s.d.]. [Consulté le 4 août 2022]. Disponible à l'adresse : https://bitcurator.net/bitcurator/

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]. Disponible à l'adresse : https://dl.acm.org/doi/10.1145/2133806.2133826

BLEI, David, NG, Andrew et JORDAN, Michael, 2001. Latent Dirichlet Allocation. *The Journal of Machine Learning Research* [en ligne]. 1 janvier 2001. Vol. 3, pp. 601-608. [Consulté le 4 août 2022]. Disponible à l'adresse : https://www.researchgate.net/publication/221620547_Latent_Dirichlet_Allocation

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

BOOMS, Hans, 1972. Gesellschaftsformen und Überlieferungsbildung Zur Problemakik Archivalischer Quellenbewertung. *Archivarische Zeitschrift*. 1972. Vol. 68, pp. 3-40.

BOOMS, Hans, 2001. Ordre social et constitution du patrimoine archivistique. *Archives* [en ligne]. 2002 2001. Vol. 33, n° 3-4, pp. 38. [Consulté le 20 mai 2022]. Disponible à l'adresse : https://www.archivistes.qc.ca/revuearchives/vol33_3-4/33-3-4-booms.pdf

BORGES, Jorge Luis, 1956. Funes ou la mémoire. In : *Fictions*. Paris : Gallimard. pp. 109-118.

BRIEL, Jutta, 2001. Workshop Bewertungspraxis « Alles nur Fingerspitzengefühl? »: Einführender Vortrag anlässlich der 2. Arbeitstagung des VKA. *Verband Schleswig-Holsteinischer Kommunalarchivarinnen und -Archivare: Mitteilungen. 2001*. pp. 48-53.

BROOKS, Philipp C., 1940. The Selection of Records for Preservation. *The American Archivist*. 1940. Vol. 3, n° 4, pp. 221-234.

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].. Disponible à l'adresse : https://doi.org/10.1080/23257962.2016.1145577

BURGY, François, EGLI, Anita et SCHMUTZ, Jürg, 2007. Évaluation et sélection des documents dans les archives suisses : éliminer avec discernement et constituer le patrimoine. In : COUTAZ, Gilbert, HUBERT, Rodolfo, KELLERHALS, Andreas, PFIFFNER, Albert et ROTH-LOCHNER, Barbara, *Archivpraxis in der Schweiz = Pratiques archivistiques en Suisse*. Baden : Hier+Jetzt. pp. 279-302.

BURGY, François et ROTH-LOCHNER, Barbara, 2002. Les Archives en Suisse ou la fureur du particularisme. *Archives* [en ligne]. 2002-2003. Vol. 34, n° 1-2, pp. 37-90. Disponible à l'adresse : https://archivistes.qc.ca/revuearchives/vol34_1-2/34-1-2-Burgy-Roth.pdf

BÜTIKOFER, Niklaus, 1995. Bewertung als Priorisierung. *arbido*. 1995. Vol. 11, pp. 14-16.

CHABIN, Marie-Anne, 2006. Cycle de vie du document engageant : Approche logistique (française) versus approche par le statut de l'information (anglo-saxonne). Wikimedia Commons [en ligne]. [Consulté le 13 juillet 2022]. Disponible à l'adresse : https://commons.wikimedia.org/wiki/File:Cycle_de_vie_document-record_mac.png?uselang=fr

CHABIN, Marie-Anne, 2007. *Archiver, et après ?* Paris : Djakarta éd. ISBN 978-2-9528828-0-4. 025.009 44

CHABIN, Marie-Anne, 2008. E-records management et diplomatique numérique. In : *Traitement et pratiques documentaires : vers un changement de paradigme ?* [en ligne]. Paris : CNAM. 2008. [Consulté le 7 juillet 2022]. Sciences et techniques de l'information. Disponible à l'adresse : https://docplayer.fr/17617420-E-records-management-et-diplomatique-numerique.html

CHABIN, Marie-Anne, 2013. Peut-on parler de diplomatique numérique ? In : FREY, Valentin et TRELEANI, Matteao, *Vers un nouvel archiviste numérique* [en ligne]. Paris : L'Harmattan.

[Consulté le 7 juillet 2022]. Disponible à l'adresse : https://www.marieannechabin.fr/diplomatique-numerique/

CHABIN, Marie-Anne et WATEL, Françoise, 2006. L'approche française du records management : concepts, acteurs et pratiques. I [en ligne]. 2006. Vol. 204, n° 4, pp. 113-130. [Consulté le 10 juillet 2022]. Disponible à l'adresse : https://www.persee.fr/docAsPDF/gazar_0016-5522_2006_num_204_4_3830.pdf

CHANEY, Allison, WALLACH, Hanna, CONNELLY, Matthew et BLEI, David, 2016. Detecting and Characterizing Events. In : *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* [en ligne]. Austin, Texas : Association for Computational Linguistics. Novembre 2016. pp. 1142-1152. [Consulté le 19 juin 2022]. Disponible à l'adresse : https://aclanthology.org/D16-1122

Checksum. *Dictionary of Archives Terminology* [en ligne]. 2022. [Consulté le 19 juin 2022] Disponible à l'adresse : https://dictionary.archivists.org/entry/checksum.html

CHEVIEUX, Caroline, 2019. *Évaluation des documents du service Données et Archives de la RTS pour une meilleure gestion de leur cycle de vie* [en ligne]. Travail de Bachelor. Genève : Haute École de Gestion Genève. [Consulté le 7 juillet 2022]. Disponible à l'adresse : https://doc.rero.ch/record/327837/files/ChevieuxCaroline_TB_memoire_version_finale.pdf

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]. Disponible à l'adresse : https://doi.org/10.1108/00012531111135628

COCCIOLO, Anthony, 2016. Finding Inactive Records on Institutional Networks: an Evaluation of Tools. *Practical Technology for Archives* [en ligne]. 16 juin 2016. [Consulté le 14 juin 2022]. Disponible à l'adresse : https://practicaltechnologyforarchives.org/issue6_cocciolo/

COLAVIZZA, Giovanni, BLANKE, Tobias, JEURGENS, Charles et NOORDEGRAAF, Julia, 2022. Archives and AI: An Overview of Current Debates and Future Perspectives. *Journal on Computing and Cultural Heritage* [en ligne]. 28 février 2022. Vol. 15, n° 1, pp. 1-15. [Consulté le 12 juin 2022]. Disponible à l'adresse : https://dl.acm.org/doi/10.1145/3479010

COLAVIZZA, Giovanni, EHRMANN, Maud et BORTOLUZZI, Fabio, 2019. Index-Driven Digitization and Indexation of Historical Archives. *Frontiers in Digital Humanities* [en ligne]. 2019. Vol. 6. [Consulté le 19 juin 2022]. Disponible à l'adresse : https://www.frontiersin.org/article/10.3389/fdigh.2019.00004

COMMUNITY ARCHIVES AND HERITAGE GROUP, 2018. *Digital Preservation for Community Archives* [en ligne]. 2018. [Consulté le 16 juin 2022]. Disponible à l'adresse : https://www.communityarchives.org.uk/wp-content/uploads/2018/02/Digital-Preservation-for-Community-Archives-V1.4-2018.pdf

CONSEIL D'ÉTAT DE NEUCHÂTEL, 2021. *Rapport de gestion financière, Tome 2 : Vision par département et entité 2020* [en ligne]. Comptes. Neuchâtel : État de Neuchâtel. [Consulté le 14 juillet 2022]. Disponible à l'adresse : https://www.ne.ch/autorites/DFS/SFIN/Documents/Comptes/C2020CE_Tome2.pdf

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.

COOK, Terry, 1994. Electronic records, paper minds: the revolution in information management and archives in the post-custodial and post-modernist era. *Archives and Manuscripts*. 1994. Vol. 22, n° 2, pp. 300-328.

COUTAZ, Gilbert, 2011. Le calendrier de conservation : Le cœur de la politique d'archivage des Archives cantonales vaudoises [en ligne]. *Rapport d'activité*. Archives cantonales vaudoises. [Consulté le 20 mai 2022]. Disponible à l'adresse : https://www.vd.ch/fileadmin/user_upload/organisation/chancellerie/ACV/fichiers_pdf/dossier-thematique/Dossier-thematique-2011.pdf

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. [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/

COUTAZ, Gilbert, HUBER, Rodolfo, KELLERHALS, Andreas, PFIFFNER, Albert et ROTH-LOCHNER, Barbara (éd.), 2007. *Archivpraxis in der Schweiz: = Pratiques archivistiques en Suisse*. Baden : hier + jetzt, Verlag für Kultur und Geschichte. ISBN 978-3-03919-045-4.

COUTURE, Carol, 1996. L'évaluation des archives. État de la question. *Archives* [en ligne]. 1997-1996. Vol. 28, n° 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, 1994. *Les fondements de la discipline archivistique*. Sainte-Foy, Québec : Presses de l'Université du Québec.

CUNNINGHAM, Adrian, 1996. Journey to the end of the night : custody and the dawning of a new era on the archival threshold. *Archives and Manuscripts*. 1996. Vol. 24, n° 2, pp. 312-321.

DEBENATH, Olivier, KAISER, Martin, KANSY, Lambert, LÜTHI, Martin et RYTER, Stefan, 2016. *Modèle conceptuel pour logiciels de gestion d'archives* [en ligne]. Document de travail. Berne : Centre de coordination pour l'archivage à long terme de documents électroniques. [Consulté le 15 mai 2022]. Disponible à l'adresse : https://kost-ceco.ch/cms/dl/929d87a28f83bbdab11ec8e78021abfa/KOSTDiskussionsPapier-AIS-Modell-v1-2_fr.pdf?target=1

DECKER, Stephanie, KIRSCH, David A., KUPPILI VENKATA, Santhilata et NIX, Adam, 2021. Finding light in dark archives: using AI to connect context and content in email. *AI & SOCIETY* [en ligne]. 31 décembre 2021. [Consulté le 13 juin 2022]. Disponible à l'adresse : https://link.springer.com/10.1007/s00146-021-01369-9

DEKENS, Charline, 2011. Quelles ressources électroniques pour quel records management ? Une perspective. *arbido* [en ligne]. 2011. Vol. 1. [Consulté le 10 juillet 2022]. Disponible à l'adresse : https://arbido.ch/fr/edition-article/2011/streifzug-durchs-web/quelles-ressources-%C3%A9lectroniques-pour-quel-records-management-une-perspective

DIGITAL CURATION CENTER, 2022. What is digital curation? *Digital Curation Center (DCC)* [en ligne]. 2022. [Consulté le 10 août 2022]. Disponible à l'adresse : https://www.dcc.ac.uk/about/digital-curation

Disk Image. *Wikipedia* : The Free Encyclopedia. [en ligne]. Dernière modification de la page le 2 août 2022 20:35. [Consulté le 10 août 2022]. Disponible à l'adresse : https://en.wikipedia.org/w/index.php?title=Disk_image&oldid=1101988045

DOCUTEAM, 2022. docuteam packer 6.3 overview. *docuteam* [en ligne]. 2022. [Consulté le 11 août 2022]. Disponible à l'adresse : https://docs.docuteam.ch/packer/6.3/en/index

DOLLAR, Charles M., 1978. Appraising Machine-Readable Records. *The American Archivist* [en ligne]. 1 octobre 1978. Vol. 41, n° 4, pp. 423-430. [Consulté le 26 juin 2022]. Disponible à l'adresse : https://meridian.allenpress.com/american-archivist/article/41/4/423/22951/Appraising-Machine-Readable-Records

Données liées. *Grand dictionnaire terminologique* [en ligne]. 2012. [Consulté le 11 août 2022]. Disponible à l'adresse : https://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=26520043

DOOM, Vincent, 2006. L'évaluation scientifique des archives, principes et stratégies. Du melon au diamant. *Gazette des archives* [en ligne]. 2006. Vol. 202, n° 2, pp. 5-43. [Consulté le 20 mai 2022]. Disponible à l'adresse : https://www.persee.fr/docAsPDF/gazar_0016-5522_2006_num_202_2_3815.pdf

DOSZKOCS, Tamas E., 1986. Natural language processing in information retrieval. *Journal of the American Society for Information Science* [en ligne]. 1986. Vol. 37, n° 4, pp. 191-196. [Consulté le 1 août 2022]. Disponible à l'adresse : https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28198607%2937%3A4%3C191%3A%3AAID-ASI3%3E3.0.CO%3B2-Y

DRAKE, Jarrett, 2015. Archiving Email at the Princeton University Archives. *Mudd Manuscript Library Blog* [en ligne]. 29 mai 2015. [Consulté le 3 juin 2022]. Disponible à l'adresse : https://blogs.princeton.edu/mudd/2015/05/archiving-email-at-the-princeton-university-archives/

DUCHARME, Daniel, 2000. *L'identification de critères d'évaluation pour les archives informatiques. Archives* [en ligne]. 2001 2000. Vol. 32, n° 2, pp. 17-32. [Consulté le 11 avril 2022]. Disponible à l'adresse : https://www.archivistes.qc.ca/revuearchives/vol32_2/32-2-ducharme.pdf

DUCHARME, Daniel et COUTURE, Carol, 1996. *L'évaluation en archivistique, évolution et tendances*. [en ligne]. 1997 1996. Vol. 28, n° 1, pp. 40. [Consulté le 19 mai 2022]. Disponible à l'adresse : https://www.archivistes.qc.ca/revuearchives/vol28_1/28-1-ducharme-couture.pdf

DUFF, Wendy M. et HARRIS, Verne, 2002. Stories and names: Archival description as narrating records and constructing meanings. *Archival Science* [en ligne]. 2002. Vol. 2, n° 3, pp. 263-285. [Consulté le 19 juin 2022]. Disponible à l'adresse : https://link.springer.com/content/pdf/10.1007/BF02435625.pdf

DUFF, Wendy M. et JOHNSON, Catherine A., 2002. Accidentally Found on Purpose: Information-Seeking Behavior of Historians in Archives. *The Library Quarterly: Information, Community, Policy* [en ligne]. 2002. Vol. 72, n° 4, pp. 472-496. [Consulté le 11 août 2022]. Disponible à l'adresse : https://www.jstor.org/stable/40039793

DUNANT GONZENBACH, Anouk, 2022. Le sort final pressenti : une tentative pragmatique d'application de l'archivistique post-moderniste? *Le présent d'hier et de demain* [en ligne]. 20 juillet 2022. [Consulté le 20 juillet 2022]. Disponible à l'adresse :

http://hieretdemain.ch/2022/07/20/le-sort-final-pressenti-une-tentative-pragmatique-dapplication-de-larchivistique-post-moderniste/

DUNANT GONZENBACH, Anouk et DUCRY, Emmanuel, 2014. L'archivage des documents électroniques à Genève, aspects organisationnels et techniques: le projet Gal@tae". In : HIRAUX, Françoise et MIRGUET, Françoise (éd.), *Actes des 13es Journées des Archives [en ligne]. Louvain-la-Neuve : Academia-Bruylant*. 2014. pp. 73-85. [Consulté le 3 juillet 2022]. Disponible à l'adresse : http://hieretdemain.gonzen.com/2013_galatae_blog.pdf

DURANTI, Luciana, 1989. Diplomatics: New Uses for an Old Science, Part II. *Archivaria* [en ligne]. 1 janvier 1989. pp. 4-17. [Consulté le 12 juillet 2022]. Disponible à l'adresse : https://archivaria.ca/index.php/archivaria/article/view/11605

DURANTI, Luciana, 1997. The Archival Bond. *Archives and Museum Informatics* [en ligne]. septembre 1997. Vol. 11, pp. 213-218. [Consulté le 9 juin 2022]. Disponible à l'adresse : https://link.springer.com/content/pdf/10.1023/A:1009025127463.pdf

DURANTI, Luciana, 1998. *Diplomatics: new uses for an old science*. Lanham, Md. : Scarecrow Press. ISBN 978-0-8108-3528-3.

DURANTI, Luciana, 2003. Pour une diplomatique des documents électroniques. *Bibliothèque de l'École des chartes* [en ligne]. 2003. Vol. 161, n° 2, pp. 603-623. [Consulté le 7 juillet 2022]. Disponible à l'adresse : https://www.jstor.org/stable/42959892

DURANTI, Luciana, 2007. The InterPARES 2 Project (2002-2007): An Overview. *Archivaria* [en ligne]. 2007. pp. 113-121. [Consulté le 3 août 2022]. Disponible à l'adresse : https://archivaria.ca/index.php/archivaria/article/view/13155

DURANTI, Luciana et CHABIN, Marie-Anne, 2004. La conservation à long terme des documents dynamiques et interactifs: InterPARES 2. *Document numérique* [en ligne]. 2004. Vol. 8, n° 2, pp. 73-86. [Consulté le 3 août 2022]. Disponible à l'adresse : https://www.cairn.info/revue-document-numerique-2004-2-page-73.htm

DURANTI, Luciana et MACNEIL, Heather, 1996. The Protection of the Integrity of Electronic Records: An Overview of the UBC-MAS Research Project. *Archivaria* [en ligne]. 1996. Vol. 42, pp. 46-67. [Consulté le 9 juin 2022]. Disponible à l'adresse : https://archivaria.ca/index.php/archivaria/article/download/12153/13158/0

EASTWOOD, Terry, 1992. Towards a social theory of appraisal archivists. In : CRAIG, Barbara L., *The Archival Imagination: Essays in Honour of Hugh A. Taylor*. Ottawa : Association of Canadian Archivists. pp. 7-89.

EASTWOOD, Terry, CRAIG, Barbara, MEI, Du, EPPARD, Philip, FIORAVANTI, Gigliola, FORTIER, Normand, GIGUERE, Mark, HANNIGAN, Ken, HORSMAN, Peter, JONKER, Agnes, STOUT, Leon et CHEN, Su-Shing, 2005. Part Two: Choosing to preserve. The selection of electronic record. In : DURANTI, Luciana (éd.), *The InterPares Project. The Long-term Preservation of Authentic Electronic Records: The Findings of the InterPARES Project* [en ligne]. San Miniato : InterPares. [Consulté le 26 juin 2022]. Disponible à l'adresse : http://www.interpares.org/book/interpares_book_e_part2.pdf

EGGER, Florian, DESJOBERT, Matthieu et BONGIOVANNI, David, 2018. V. 1.0 : *ArchiSelect Persona & Scénarios*. Genève : Telono S.A. Document interne

EHRMANN, Maud, HAMDI, Ahmed, PONTES, Elvys Linhares, ROMANELLO, Matteo et DOUCET, Antoine, 2021. Named Entity Recognition and Classification on Historical Documents: A Survey. *arXiv:2109.11406 [cs]* [en ligne]. 23 septembre 2021. pp. 39. [Consulté le 10 décembre 2021]. Disponible à l'adresse : http://arxiv.org/abs/2109.11406

ELINGS, Mary W., 2016. Using NLP to Support Dynamic Arrangement, Description, and Discovery of Born Digital Collections: The ArchExtract Experiment. *bloggERS!* [en ligne]. 24 mai 2016. [Consulté le 23 novembre 2021]. Disponible à l'adresse : https://saaers.wordpress.com/2016/05/24/using-nlp-to-support-dynamic-arrangement-description-and-discovery-of-born-digital-collections-the-archextract-experiment/

ELRAGAL, Ahmed et PÄIVÄRINTA, Tero, 2017. Opening Digital Archives and Collections with Emerging Data Analytics Technology: A Research Agenda. *Tidsskriftet Arkiv* [en ligne]. 19 février 2017. Vol. 8, n° 1. [Consulté le 21 novembre 2021]. Disponible à l'adresse : https://journals.oslomet.no/index.php/arkiv/article/view/1959

ESPOSITO, F., MALERBA, D., SEMERARO, G., FERILLI, S., ALTAMURA, O., BASILE, T.M.A., BERARDI, M., CECI, M. et DI MAURO, N., 2004. Machine learning methods for automatically processing historical documents: from paper acquisition to XML transformation. In : *First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings*. [en ligne]. Palo Alto, CA, USA : IEEE. 2004. pp. 328-335. [Consulté le 5 août 2022]. ISBN 978-0-7695-2088-9. Disponible à l'adresse : http://ieeexplore.ieee.org/document/1263262/

FAVIER, Jean et NEIRINCK, Daniel (éd.), 1993. *La pratique archivistique française*. Paris : Archives Nationales. ISBN 978-2-86000-205-9.

FIORUCCI, Marco, KHOROSHILTSEVA, Marina, PONTIL, Massimiliano, TRAVIGLIA, Arianna, DEL BUE, Alessio et JAMES, Stuart, 2020. Machine Learning for Cultural Heritage: A Survey. *Pattern Recognition Letters* [en ligne]. 1 mai 2020. Vol. 133, pp. 102-108. [Consulté le 19 juin 2022]. Disponible à l'adresse : https://www.sciencedirect.com/science/article/pii/S0167865520300532

Fingerspitzengefühl. *Leo*. [en ligne]. 2022. [Consulté le 12 juillet 2022] Disponible à l'adresse : https://dict.leo.org/allemand-fran%C3%A7ais/Fingerspitzengef%C3%BChl

FOLDERMATCH, 2020. *FolderMatch v4 - Compare Files demo* [enregistrement vidéo]. *YouTube* [en ligne]. 6 mars 2020. [Consulté le 3 août 2022]. Disponible à l'adresse : https://www.youtube.com/watch?v=oRfl49WxejA

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.

FRITZ, Leonie, 2021. *Tools zur Übernahme digitaler Dateiablagen : Ein Test von Archifiltre, DROID und TreeSize Professional zur Umsetzung der technischen Analyse von Privatnachlässen anhand des Musterworkflows der KOST-Arbeitsgruppe «Dateiablage»*. CAS ALIS Zertifikatarbeit. Lausanne; Berne : Université de Lausanne; Universität Bern.

GAUDINAT, Arnaud, 2016. Le plaisir de tout conserver sans modération : une question de taille ?. *arbido* [en ligne]. Mars 2016. [Consulté le 13 mai 2021]. Disponible à l'adresse : https://arbido.ch/fr/edition-article/2016/d%C3%A9truire-pour-conserver/le-plaisir-de-tout-conserver-sans-mod%C3%A9ration-une-question-de-taillehttps://arbido.ch/fr/

GAUDINAT, Arnaud et KNAFOU, Julien, 2017a. *Modèle d'évaluation systématique des documents : Concept & Preuve de concept, Axe 2, Livrable 1 : Mandat d'accompagnement « fouille de données » : État de l'art*. Rapport de recherche. Genève : Haute École de Gestion Genève. Document interne

GAUDINAT, Arnaud et KNAFOU, Julien, 2017b. *Modèle d'évaluation systématique des documents : Concept & Preuve de concept, Axe 2, Livrable 2 : Mandat d'accompagnement «*

fouille de données » : Description des données et propositions d'indicateurs clés possibles. Rapport de recherche. Genève : Haute École de Gestion Genève. Document interne

GAUDINAT, Arnaud et KNAFOU, Julien, 2018a. *Modèle d'évaluation systématique des documents : Concept & Preuve de concept, Axe 2, Livrable 6 : Mandat d'accompagnement « fouille de données » : Spécification, fouille de donnée - Rapport final*. Rapport de recherche. Genève : Haute École de Gestion Genève. Document interne

GAUDINAT, Arnaud et KNAFOU, Julien, 2018b. *Modèle d'évaluation systématique des documents : Concept & Preuve de concept, Axe 2, Livrable 3 : Mandat d'accompagnement « fouille de données » : Analyse des métadonnées externes aux dossiers d'activité*. Rapport de recherche. Genève : Haute École de Gestion Genève. Document interne à l'institution

GAUDINAT, Arnaud et KNAFOU, Julien, 2018c. *Modèle d'évaluation systématique des documents : Concept & Preuve de concept, Axe 2, Livrable 4 : Mandat d'accompagnement « fouille de données » : Preuve de concept de fouille de données*. Rapport de recherche. Genève : Haute École de Gestion Genève. Document interne

GAUDINAT, Arnaud et KNAFOU, Julien, 2018d. *Modèle d'évaluation systématique des documents : Concept & Preuve de concept, Axe 2, Livrable 5 : Mandat d'accompagnement « fouille de données » : Test de faisabilité du modèle opérationnel*. Rapport de recherche. Genève : Haute École de Gestion Genève. Document interne

GAUDINAT, Arnaud et KNAFOU, Julien, 2018e. *Modèle d'évaluation systématique des documents : Concept & Preuve de concept, Axe 2, Livrable 5.3 : Mandat d'accompagnement « fouille de données » : Résultat de l'enquête auprès des archivistes.* Rapport de recherche. Genève : Haute École de Gestion Genève. Document interne

GAUDINAT, Arnaud et KNAFOU, Julien, 2018f. *Modèle d'évaluation systématique des documents : Concept & Preuve de concept, Axe 2, Annexes Livrable 6*. Rapport de recherche. Genève : Haute École de Gestion Genève. Document interne

GILLILAND, Anne J., 2014. *Conceptualizing 21st-century archives*. Chicago : Society of American Archivists. ISBN 978-1-931666-68-8.

GILLILAND, Anne J., 2016. Designing Expert Systems for Archival Evaluation and Processing of Computer Mediated Communications: Frameworks and Methods. In : GILLILAND, Anne J., MCKEMMISH, S. et LAU, A. J., *Research in the Archival Multiverse* [en ligne]. Melbourne : Monash University Press. pp. 685-721. [Consulté le 13 juin 2022]. Disponible à l'adresse : https://escholarship.org/uc/item/6c21p3hs

GOLLINS, Timothy, MCDONALD, Graham, MACDONALD, Craig et OUNIS, Iadh, 2014. On Using Information Retrieval for the Selection and Sensitivity Review of Digital Public Records. In : *Proceeding of the 1st International Workshop on Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security* [en ligne]. Gold Coast, Australia. 11 juillet 2014. [Consulté le 22 novembre 2021]. Disponible à l'adresse : http://ceur-ws.org/Vol-1225/pir2014_submission_9.pdf

GOODFELLOW, Ian, BENGIO, Yoshua et COURVILLE, Aaron, 2016. *Deep Learning* [en ligne]. MIT Press. [Consulté le 4 août 2022]. Disponible à l'adresse : https://www.deeplearningbook.org/

GOODING, Paul, 2016. Exploring the information behaviour of users of Welsh Newspapers Online through web log analysis. *Journal of Documentation* [en ligne]. 1 janvier 2016. Vol. 72, n° 2, pp. 232-246. [Consulté le 11 août 2022]. Disponible à l'adresse : https://doi.org/10.1108/JD-10-2014-0149

GOODMAN, Morgan M., 2019. *« What Is On This Disk? » An Exploration Of Natural Language Processiong In Archival Appraisal* [en ligne]. Master's Paper for the M.S. in IS degree. Chapel Hill : School of Information and Library Science of the University of North Carolina. [Consulté le 25 mai 2022]. Disponible à l'adresse : https://cdr.lib.unc.edu/downloads/wm117s91s?locale=en

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].. Disponible à l'adresse : https://doi.org/10.17723/aarc.68.2.c741823776k65863

HAENER, Ruth, 1995. Von Qualität zur Quantität : Einführung in die Diskussion der Bewertungstheorie. *arbido*. 1995. Vol. 9, pp. 15-18.

HALBEISEN, Patrick, 1999. *Von der vorarchivischen Schriftgutverwaltung zur vorarchivischen Bewertung: Konzeptionelle Überlegungen zum Aufbau eines Bankarchivs am Beispiel der Schweizerischen Kredit- anstalt: Ein Beitrag zur Bewertungsdiskussion in der Archivistik*. Bern; Stuttgart : Haupt. 178 p.

HAM, Gerald, 1981. Archival strategies for the post-custodial era. *The American Archivist*.1981. Vol. 44, n° 3, pp. 207-216.

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]. Disponible à l'adresse : https://www.researchgate.net/publication/220364258_Automating_the_appraisal_of_digital_materials

HAWKINS, Ashleigh, 2021. Archives, linked data and the digital humanities: increasing access to digitised and born-digital archives via the semantic web. *Archival Science* [en ligne]. 27 décembre 2021. [Consulté le 14 juin 2022]. Disponible à l'adresse : https://link.springer.com/10.1007/s10502-021-09381-0

HENGCHEN, Simon, COECKELBERGS, Mathias, VAN HOOLAND, Seth, VERBORGH, Ruben et STEINER, Thomas, 2016. Exploring archives with probabilistic models: Topic modelling for the valorisation of digitised archives of the European Commission. In : *2016 IEEE International Conference on Big Data (Big Data)* [en ligne]. Washington, DC : IEEE. décembre 2016. pp. 3245-3249. [Consulté le 2 janvier 2022]. ISBN 978-1-4673-9005-7. Disponible à l'adresse : http://ieeexplore.ieee.org/document/7840981/

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. 78-90. [Consulté le 23 juin 2022]. Disponible à l'adresse : https://arxivers.com/wp-content/uploads/2018/10/1.4_-Dossier_SVHooland_MCoeckelbergs.pdf

HUBER, 2009. Archivische Bewertung: Aspekte, Probleme, Konjunkturen. *arbido* [en ligne]. 2009. N° 4. [Consulté le 26 juillet 2022]. Disponible à l'adresse : https://www.arbido.ch/de/ausgaben-artikel/2009/bewertung-als-kernaufgabe-der-i-d-welt/archivische-bewertung-aspekte-probleme-konjunkturen

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]. Disponible à l'adresse : https://www.emerald.com/insight/content/doi/10.1108/RMJ-09-2019-0055/full/html

INTERPARES TRUST AI, 2021a. InterPARES Trust AI - Artificial Intelligence. *ITrust AI* [en ligne]. 2021. [Consulté le 3 août 2022]. Disponible à l'adresse : https://interparestrustai.org/

INTERPARES TRUST AI, 2021b. Research Studies. *ITrust AI* [en ligne]. 2021. [Consulté le 3 août 2022]. Disponible à l'adresse : https://interparestrustai.org/trust/about_research/studies

JAILLANT, Lise et CAPUTO, Annalina, 2022. Unlocking digital archives: cross-disciplinary perspectives on AI and born-digital data. *AI & SOCIETY* [en ligne]. 12 janvier 2022. [Consulté le 8 juillet 2022]. Disponible à l'adresse : https://doi.org/10.1007/s00146-021-01367-x

JAIN, Anil K., 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* [en ligne]. juin 2010. Vol. 31, n° 8, pp. 651-666. [Consulté le 20 juin 2022]. Disponible à l'adresse : https://linkinghub.elsevier.com/retrieve/pii/S0167865509002323

JAM SOFTWARE, [s.d]. Remove file duplicates with TreeSize. *Jam Software* [en ligne]. [s.d]. [Consulté le 3 août 2022]. Disponible à l'adresse : https://www.jam-software.com/treesize/deduplicate_files.shtml

JENKINSON, Hilary, 1937. *A Manual of Archive Administration Including the Problems of War Archives and Archive Making*. 2nd. Oxford : The Clarendon Press.

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, REDWINE, Gabriela et DONAHUE, Rachel, 2010. *Digital forensics and born-digital content in cultural heritage collections* [en ligne]. Washington, D.C : Council on Library and Information Resources. [Consulté le 1 août 2022]. CLIR publication, 149. ISBN 978-1-932326-37-6. Disponible à l'adresse : https://www.clir.org/wp-content/uploads/sites/6/pub149.pdf

KNOBLOCH, Corinna, 2019. Archivischer Umgang mit digitalen Sammlungen am Beispiel der Johannes-Wagner-Schule Nürtingen. In : STUMPF, Marcus et TIEMANN, Katharina, *Erziehung und Bildung als kommunalarchivische Überlieferungsfelder. Beiträge des 27. Fortbildungsseminars der Bundeskonferenz der Kommunalarchive (BKK)* [en ligne]. Münster. pp. 76-86. [Consulté le 26 juin 2022]. ISBN 978-3-936258-29-5. Disponible à l'adresse : https://www.lwl-archivamt.de/media/filer_public/54/fd/54fda6a4-23b6-4fb9-86de-d54ed7984c0f/tua_35_P6OTz4x.pdf

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

LEE, Christopher A., CHASSANOFF, Alexandra, WOODS, Kam, KIRSCHENBAUM, Matthew et OLSEN, Porter, 2012. BitCurator: Tools and Techniques for Digital Forensics in Collecting Institutions. *D-Lib Magazine* [en ligne]. Mai 2012. Vol. 18, n° 5/6. [Consulté le 24 juillet 2022]. Disponible à l'adresse : http://www.dlib.org/dlib/may12/lee/05lee.html

LEE, Christopher A. et WOODS, Kam, 2017. Diverse digital collections meet diverse uses: applying natural language processing to born-digital primary sources. *iPres 2017: the 14th International Conference on Digital Preservation* [en ligne]. 2017. [Consulté le 12 décembre

2021]. Disponible à l'adresse : https://ipres-conference.org/ipres17/ipres2017.jp/wp-content/uploads/50.pdf

LEMAY, Yvon et KLEIN, Anne, 2014. Les archives définitives : un début de parcours. Revisiter le cycle de vie et le Records continuum. *Archivaria*. 2014. N° 77, pp. 73-102.

LENARTZ, Stephan, 2020. *Digital ist Besser? : Möglichkeiten der automatisierten Aufbereitung und Bewertung von Fileablagen mit Python am Beispiel einer digitalen Fotosammlung* [en ligne]. Landesarchiv Baden-Württemberg. Stuttgart. [Consulté le 25 juin 2022]. Disponible à l'adresse : https://www.landesarchiv-bw.de/media/full/70717

Loi fédérale sur l'archivage du 26 juin 1998 (LAr ; 152.1). *Assemblée fédérale de la Confédération suisse* [en ligne]. 26 juin 1998. Mise à jour le 1er mai 2013. [Consulté le 15 juillet 2022] Disponible à l'adresse : https://www.fedlex.admin.ch/eli/cc/1999/354/fr

Loi sur l'archivage du 22 février 2011 (LArch ; 442.20). *Le Grand Conseil de la République et Canton de Neuchâtel* [en ligne]. 22 février 2011. Mise à jour le 1er janvier 2021. [Consulté le 9 juin 2022] Disponible à l'adresse : https://rsn.ne.ch/DATA/program/books/rsne/htm/44220.htm

Loi sur les archives de l'État du 9 octobre 1989 (442.20). *Le Grand Conseil de la République et Canton de Neuchâtel* [en ligne]. [Consulté le 14 juillet 2022] Disponible à l'adresse : https://rsn.ne.ch/DATA/program/books/RSN2010/20091/htm/44220.htm

LUTKEVICH, Ben et BURNS, Ed, 2021. What is Natural Language Processing? An Introduction to NLP. *SearchEnterpriseAI* [en ligne]. 2021. [Consulté le 2 décembre 2021]. Disponible à l'adresse : https://searchenterpriseai.techtarget.com/definition/natural-language-processing-NLP

MAKHLOUF SHABOU, Basma, 2010. *Étude sur la définition et la mesure des qualités des archives définitives issues d'une évaluation* [en ligne]. Thèse de doctorat. Montréal : Université de Montréal. [Consulté le 25 mars 2021]. Disponible à l'adresse : https://papyrus.bib.umontreal.ca/xmlui/bitstream/handle/1866/4955/Makhouf-Shabou_Basma_2011_these.pdf?sequence=5&isAllowed=y

MAKHLOUF SHABOU, Basma, 2012a. *Comment évaluer la qualité des archives ? : Méthode et instruments de mesure des dimensions de qualité des archives définitives*. Saarbrücken : Éditions Universitaires Européennes. ISBN 978-3-8417-9514-4.

MAKHLOUF SHABOU, Basma, 2012b. Étude sur la définition et la mesure des qualités des archives définitives issues d'une évaluation. *Archives* [en ligne]. 1 janvier 2012. Vol. 43, n° 2, pp. 39-70. [Consulté le 5 juillet 2022]. Disponible à l'adresse : https://www.archivistes.qc.ca/revuearchives/vol43_2/43_2_makhlouf-shabou.pdf

MAKHLOUF SHABOU, Basma, 2013. *QADEPs : Définition et mesure des qualités des archives et documents électroniques publics*. Rapport scientifique. Genève : Haute École de Gestion Genève.

MAKHLOUF SHABOU, Basma, 2015a. Digital diplomatics and measurement of electronic public data qualities: What lessons should be learned? LUCIANA DURANTI, Dr (éd.), Records *Management Journal* [en ligne]. 1 janvier 2015. Vol. 25, n° 1, pp. 56-77. [Consulté le 25 mars 2021]. Disponible à l'adresse : https://doi.org/10.1108/RMJ-01-2015-0006

MAKHLOUF SHABOU, Basma, 2015b. 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. Québec : Presses de l'Université du Québec. pp. 195-218.

MAKHLOUF SHABOU, Basma et TIÈCHE, Julien, 2018a. *Métriques archivistiques pour l'évaluation, Livrables 1.1 & 1.2 : Cadre conceptuel et typologie des critères*. Rapport de recherche. Genève : Haute École de Gestion Genève. Document interne

MAKHLOUF SHABOU, Basma et TIÈCHE, Julien, 2018b. *Métriques archivistiques pour l'évaluation, Livrables 2.1 & 2.2 : Modèle opérationnel détaillé des métriques archivistiques et Nomenclature des critères et métriques d'évaluation*. Rapport de recherche. Genève : Haute École de Gestion Genève. Document interne

MAKHLOUF SHABOU, Basma et TIÈCHE, Julien, 2018c. *Métriques archivistiques pour l'évaluation, Livrable 3.1 : Études de cas*. Rapport de recherche. Genève : Haute École de Gestion Genève. Document interne

MAKHLOUF SHABOU, Basma et TIÈCHE, Julien, 2018d. *Concept et preuve de concept de métriques pour l'évaluation archivistique, Livrable 3.2 : Définition et validation de sets de métriques : Rapport de recherche*. Rapport de recherche. Genève : Haute École de Gestion Genève. Document interne

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]. Disponible à l'adresse : https://www.emerald.com/insight/content/doi/10.1108/RMJ-09-2019-0049/full/html

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

MARSDEN, Paul, 1997. When is the future ? : comparative notes on the electronic record-keeping projects of the University of Pittsburgh and the University of British Columbia. *Archivaria*. 1997. Vol. 43, pp. 158-173.

MCCORDUCK, Pamela, 2018. *Machines who think: a personal inquiry into the history and prospects of artificial intelligence*. Boca Raton London New York : CRC Press. An A K Peters book. ISBN 978-1-56881-205-2.

MCDONALD, Graham, MACDONALD, Craig et OUNIS, Iadh, 2017. Enhancing Sensitivity Classification with Semantic Features Using Word Embeddings. In : JOSE, Joemon M, HAUFF, Claudia, ALTINGOVDE, Ismail Sengor, SONG, Dawei, ALBAKOUR, Dyaa, WATT, Stuart et TAIT, John (éd.), *Advances in Information Retrieval* [en ligne]. Cham : Springer International Publishing. pp. 450-463. Lecture Notes in Computer Science. [Consulté le 20 juin 2022]. ISBN 978-3-319-56607-8. Disponible à l'adresse : http://link.springer.com/10.1007/978-3-319-56608-5_35

MCKEMMISH, Susan Marilyn, UPWARD, Franklyn Herbert et REED, Barbara, 2010. Records continuum model. BATES, Marcia J. et MAACK, Mary Niles (éd.), *Encyclopedia of Library and Information Sciences* [en ligne]. Third Edition. London : Taylor & Francis. [Consulté le 19 juin 2022]. Disponible à l'adresse : https://research.monash.edu/en/publications/records-continuum-model

MELLIFLUO, Laure, 2008. *Évaluation des archives : en théorie et en pratique aux Archives communales de la Ville de Carouge* [en ligne]. Travail de Bachelor. Genève : Haute École de Gestion Genève. [Consulté le 19 mai 2022]. Disponible à l'adresse : https://doc.rero.ch/record/11281/files/Travail_de_bachelor.pdf

Mise en correspondance. *Grand dictionnaire terminologique* [en ligne]. 2012. [Consulté le 6 juin 2022]. Disponible à l'adresse : https://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=8873847

MITCHELL, Tom, 1997. *Machine Learning textbook* [en ligne]. New York : McGraw-Hill. [Consulté le 25 juillet 2022]. ISBN 0-07-042807-7. Disponible à l'adresse : http://www.cs.cmu.edu/~tom/mlbook.html

MORETTI, Franco, 2013. *Distant Reading*. London : Verso. ISBN 978-1-78168-084-1.

MOSS, Michael et GOLLINS, Tim, 2017. Our Digital Legacy: an Archival Perspective. *Journal of Contemporary Archival Studies* [en ligne]. 8 décembre 2017. Vol. 4, n° 2. [Consulté le 19 juin 2022]. Disponible à l'adresse : https://elischolar.library.yale.edu/jcas/vol4/iss2/3

MOSS, Michael, THOMAS, David et GOLLINS, Tim, 2018. The Reconfiguration of the Archive as Data to Be Mined. *Archivaria* [en ligne]. 26 novembre 2018. Vol. 86, pp. 118-151. [Consulté le 19 juin 2022]. Disponible à l'adresse : https://archivaria.ca/index.php/archivaria/article/view/13646

MULLER, Samuel, FEITH, Johan Adriaan et FRUIN, Robert, 1910. *Manuel pour le classement et la description des archives (adapté et traduit par J. Cuvelier et H. Stein)*. La Haye : A. de Jager.

MUMMA, Courtney C., DINGWALL, Glenn et BIGELOW, Sue, 2011. A First Look at the Acquisition and Appraisal of the 2010 Olympic and Paralympic Winter Games Fonds: or, SELECT * FROM VANOC _Records AS Archives WHERE Value="true". *Archivaria* [en ligne]. 2 décembre 2011. Vol. 72, pp. 93-122. Disponible à l'adresse : https://archivaria.ca/index.php/archivaria/article/view/13361

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

NAUGHLER, Harold, 1983. *Appraisal of machine readable records : a RAMP study with guidelines*. Paris : UNESCO.

NAUMANN, Kai, 2017. Wie es mit der Projektsammlung von Susanne Belovari weiterging. In : PUCHTA, Michael et NAUMANN, Kai, *Kreative digitale Ablagen und die Archive: Ergebnisse eines Workshops des KLA-Ausschusses Digitale Archive am 22./23.11.2016 in der Generaldirektion der Staatlichen Archive Bayerns*. München : Generaldirektion der staatlichen Archive Bayerns. pp. 30-31.

NESMITH, Tom, 2002. Seeing Archives: Postmodernism and the Changing Intellectual Place of Archives. *The American Archivist* [en ligne]. 2002. Vol. 65, n° 1, pp. 24-41. [Consulté le 31 mai 2022]. Disponible à l'adresse : https://www.jstor.org/stable/40294187

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-sautomatisehttps://arbido.ch/fr/

O'NEILL ADAMS, Margaret, 2007. Analyzing archives and finding facts: use and users of digital data records. *Archival Science* [en ligne]. 1 mars 2007. Vol. 7, n° 1, pp. 21-36. [Consulté le 4 août 2022].. Disponible à l'adresse : https://doi.org/10.1007/s10502-007-9056-4

Open source. *InterPares Trust* [en ligne]. 2018. [Consulté le 10 août 2022]. Disponible à l'adresse : https://interparestrustai.org/terminology/term/open%20source

ORGANISATION INTERNATIONALE DE NORMALISATION, 2001. Information et documentation – « Records management » – Partie 2 : Guide Pratique. Genève : ISO, 15 septembre 2001. ISO 15489-1.

ORGANISATION INTERNATIONALE DE NORMALISATION, 2008. Information et documentation – Analyse des processus pour la gestion des informations et documents d'activité. Genève : ISO, 15 juin 2008. ISO/TR 26122.

ORGANISATION INTERNATIONALE DE NORMALISATION, 2012. Space data and information transfer systems – Open archival information system (OAIS) – Reference model. Genève : ISO, 1er septembre 2012. ISO 14721.

ORGANISATION INTERNATIONALE DE NORMALISATION, 2013. Information technology — Object Management Group Business Process Model and Notation. Genève : ISO, 1er juillet 2010. ISO/IEC 19510.

ORGANISATION INTERNATIONALE DE NORMALISATION, 2016. Information et documentation – Gestion des documents d'activité – Partie 1 : Concepts et principes. Genève : ISO, 15 avril 2016. ISO 15489-1.

O'SHAE, Greg et ALLEN, David, 1996. Living in a digital world: recognising the electronic and post-custodial realities. *Archives and Manuscripts*. 1996. Vol. 24, n° 2, pp. 286-311.

PARAPADAKIS, George, 2013. A clouded view of Records and Auto-Classification. *For what it's worth...* [en ligne]. 26 juin 2013. [Consulté le 13 juin 2022]. Disponible à l'adresse : https://4most.wordpress.com/2013/06/26/clouded-view-of-records-and-classification/

PEGDEN, Caroline, 2016. From digital dark age to digital enlightenment. *The National Archives UK blog* [en ligne]. 17 février 2016. [Consulté le 14 juin 2022]. Disponible à l'adresse : https://blog.nationalarchives.gov.uk/digital-dark-age-digital-enlightenment/

PÉROTIN, Yves, 1965. Le grenier de l'Histoire et les récoltes excédentaires. *Gazette des archives* [en ligne]. 1965. Vol. 50, n° 1, pp. 131-143. [Consulté le 27 juillet 2022]. Disponible à l'adresse : https://www.persee.fr/doc/gazar_0016-5522_1965_num_50_1_1836

PUCHTA, Michael et NAUMANN, Kai (éd.), 2017. *Kreative digitale Ablagen und die Archive: Ergebnisse eines Workshops des KLA-Ausschusses Digitale Archive am 22./23.11.2016 in der Generaldirektion der Staatlichen Archive Bayerns*. 1. Auflage. München : Generaldirektion der staatlichen Archive Bayerns. Sonderveröffentlichungen der Staatlichen Archive Bayerns, Nr. 13. ISBN 978-3-938831-81-6.

RAJOTTE, David, 2010. La réflexion archivistique à l'ère du document numérique : un bilan historique. *Archives* [en ligne]. 2011 2010. Vol. 42, n° 2, pp. 69-105. [Consulté le 5 juillet 2022]. Disponible à l'adresse : http://www.archivistes.qc.ca/revuearchives/vol42_2/42_2_rajotte.pdf

Règlement d'exécution de la loi sur l'archivage (442.23). *Le Grand Conseil de la République et Canton de Neuchâtel* [en ligne]. 29 avril 2013. Mise à jour le 25 mai 2021. [Consulté le 9 juin 2022] Disponible à l'adresse : https://rsn.ne.ch/DATA/program/books/rsne/htm/44223.htm

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]. Disponible à l'adresse : https://doi.org/10.1080/01576895.2018.1502088

ROSS, Seamus et GOW, Ann, 1999. *Digital Archaeology: Rescuing Neglected and Damaged Data Resources. A JISC/NPO Study within the Electronic Libraries (eLib) Programme on the Preservation of Electronic Materials*. [en ligne]. [Consulté le 1 août 2022]. ISBN 978-1-900508-51-3. Disponible à l'adresse : https://www.researchgate.net/publication/31869567_Digital_Archaeology_Rescuing_Neglected_and_Damaged_Data_Resources_A_JISCNPO_Study_within_the_Electronic_Libraries_eLib_Programme_on_the_Preservation_of_Electronic_Materials

ROTH-LOCHNER, Barbara, 1995. Évaluation et tri aux Archives d'État de Genève. *arbido*. 1995. Vol. 11, pp. 20-24.

SAMUELS, Helen Willa, 1986a. Who Controls the Past. *The American Archivist* [en ligne]. 1986. Vol. 49, n° 2, pp. 109-124. [Consulté le 30 mai 2022]. Disponible à l'adresse : https://www.jstor.org/stable/40292980

SAMUELS, Helen Willa, 1986b. Selecting from the past for the future: towards a strategy for historical documentation. *Proceedings of the Society of Southwest Archivists Meeting (13th, 1985, San Antonio, Texas)*. 1986. pp. 8-13.

SCHELLENBERG, Theodore R., 1956. *Modern Archives: Principles and Techniques*. Chicago : Society of American Archivists.

SCHELLENBERG, Theodore R., 1965. *Management of Archives*. Washington, DC : National Archives and Records Administration.

SCHLUDI, Ulrich, 2013. Zwischen Records Management und digitaler Archivierung. Das Dateisystem als Basis von Schriftgugverwaltung und Überlieferungsbildung. In : NAUMANN, Kai et MÜLLER, Peter, *Das neue Handwerk – Digitales Arbeiten in kleinen und mittleren Archiven. Vorträge des 72. Südwestdeutschen Archivtags am 22. und 23. Juni 2012 in Bad Bergzabern*. Stuttgart : Verlag W. Kohlhammer. pp. 20-39. ISBN 978-3-17-023091-0.

SCHMIDT, Benjamin M., 2012. Words Alone: Dismantling Topic Models in the Humanities. *Journal of Digital Humanities* [en ligne]. 2012. Vol. 2, n° 1. [Consulté le 9 décembre 2021]. Disponible à l'adresse : http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/

SCHNEIDER, J., ADAMS, C., DEBAUCHE, S., ECHOLS, R., MCKEAN, C., MORAN, J. et WAUGH, D., 2019. Appraising, processing, and providing access to email in contemporary literary archives. *Archives and Manuscripts* [en ligne]. 2 septembre 2019. Vol. 47, n° 3, pp. 305-326. [Consulté le 19 juin 2022]. Disponible à l'adresse : https://www.tandfonline.com/doi/full/10.1080/01576895.2019.1622138

SCOOTER SOFTWARE, 2022. Intelligent Comparison. *Scooter Software: Home of Beyond Compare* [en ligne]. 2022. [Consulté le 3 août 2022]. Disponible à l'adresse : https://www.scootersoftware.com/features.php

SHALLCROSS, Michael, 2015. The Work of Appraisal in the Age of Digital Reproduction. *Bentley Historical Library Curation Team Blog* [en ligne]. 15 juin 2015. [Consulté le 12 juin 2022]. Disponible à l'adresse : http://archival-integration.blogspot.com/2015/06/the-work-of-appraisal-in-age-of-digital.html

SHALLCROSS, Michael, 2016a. The End is Just a New Beginning! *Bentley Historical Library Curation Team Blog* [en ligne]. 4 novembre 2016. [Consulté le 28 juin 2022]. Disponible à l'adresse : http://archival-integration.blogspot.com/2016/11/the-end-is-just-new-beginning.html

SHALLCROSS, Michael, 2016. Appraising digital archives with Archivematica. In: *2016 IEEE International Conference on Big Data (Big Data)*. [en ligne]. Décembre 2016. pp. 3272 3276. [Consulté le 15 août 2022]. Disponible à l'adresse : https://ieeexplore.ieee.org/document/7840985

SHALLCROSS, Michael et DEROMEDI, Nancy, 2012. Automated Digital Processing at the Bentley Historical Library. *iPres 2012* [en ligne]. 2012. pp. 2. [Consulté le 28 juin 2022]. Disponible à l'adresse : https://deepblue.lib.umich.edu/bitstream/handle/2027.42/95923/BHL_iPRESpaper_20120828.pdf?sequence=1&isAllowed=y

SHARMA, Akshay, 2021. Named Entity Recognition using SpaCy (NER). *The HumAIn Blog* [en ligne]. 25 mars 2021. [Consulté le 11 août 2022]. Disponible à l'adresse : https://medium.com/in-pursuit-of-artificial-intelligence/named-entity-recognition-using-spacy-ner-da6eebd3d08

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. [Consulté le 4 août 2022]. Disponible à l'adresse : https://digitalcommons.usu.edu/westernarchives/vol5/iss1/1

SIEVERT, Carson et SHIRLEY, Kenneth, 2014. LDAvis: A method for visualizing and interpreting topics. In : *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* [en ligne]. Baltimore, Maryland, USA : Association for Computational Linguistics. 2014. pp. 63-70. [Consulté le 24 juillet 2022]. Disponible à l'adresse : http://aclweb.org/anthology/W14-3110

SLOYAN, Victoria, 2016. Born-digital archives at the Wellcome Library: appraisal and sensitivity review of two hard drives. *Archives and Records* [en ligne]. 2 janvier 2016. Vol. 37, n° 1, pp. 20-36. [Consulté le 28 juin 2022]. Disponible à l'adresse : https://www.tandfonline.com/doi/full/10.1080/23257962.2016.1144504

SPENCER, Ross, 2017. Binary trees? Automatically identifying the links between born-digital records. *Archives and Manuscripts* [en ligne]. 4 mai 2017. Vol. 45, n° 2, pp. 77-99. [Consulté le 11 novembre 2021]. Disponible à l'adresse : https://doi.org/10.1080/01576895.2017.1330158

SUN, Xiaofei, YANG, Diyi, LI, Xiaoya, ZHANG, Tianwei, MENG, Yuxian, QIU, Han, WANG, Guoyin, HOVY, Eduard et LI, Jiwei, 2021. Interpreting Deep Learning Models in Natural Language Processing: A Review. *arXiv:2110.10470[cs.CL]* [en ligne]. 2021. [Consulté le 21 juin 2022]. Disponible à l'adresse : https://arxiv.org/abs/2110.10470

TAYLOR, Isabel, 2016. *Eine hydraartige Matrjoschka: Wie wir die Fileablage eines staatlichen Schulamtes bewertet und erschlossen haben*. [en ligne]. Potsdam. 2016. [Consulté le 28 juin 2022]. Disponible à l'adresse : https://www.sg.ch/content/dam/sgch/kultur/staatsarchiv/auds-2016/archivierung-von-unterlagen-mit-besonderen-strukturen/01_TAYLOR_Vortragsfolien%20(29.02.16).pdf

Text mining. *InterPares Trust* [en ligne]. 2018. [Consulté le 10 août 2022]. Disponible à l'adresse : https://interparestrustai.org/terminology/term/text%20mining

THE NATIONAL ARCHIVES UK, 2016. *The application of technology-assisted review to born-digital records transfer, Inquiries and beyond* [en ligne]. Research report. The National Archives UK. [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, 2020. *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, 2021a. *Using AI for Digital Records Selection in Government: Guidance for records managers based on an evaluation of current marketplace solutions* [en ligne]. The National Archives UK. [Consulté le 14 juin 2022]. Disponible à l'adresse : https://cdn.nationalarchives.gov.uk/documents/using-ai-digital-selection-in-government.pdf

THE NATIONAL ARCHIVES UK, 2021b. *Using AI for digital selection in government. The National Archives - Homepage* [en ligne]. 2021. [Consulté le 18 novembre 2021]. Disponible à l'adresse : https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/research-collaboration/using-ai-for-digital-selection-in-government/

THIBODEAU, Kenneth, 1998. L'impact des archives électroniques sur la fonction d'évaluation. In : *L'évaluation des archives: des nécessités de la gestion aux exigences du témoignage, 27 mars 1998. Montréal, 3e Symposium du Groupe interdisciplinaire de recherche en archivistique (GIRA)* [en ligne]. Montréal : Université de Montréal. pp. 89-96. [Consulté le 30 mai 2022]. Disponible à l'adresse : http://gira-archives.org/files/2014/11/gira_1998.pdf

TIÈCHE, Julien, 2015. *La mesure des dimensions de la qualité des archives électroniques : apport des textes normatifs en matière d'archivage électroniques : apport des textes normatifs en matière d'archivage électronique à long terme* [en ligne]. Travail de Bachelor HES. Genève : Haute École de Gestion Genève (HEG-GE). [Consulté le 30 mai 2022]. Disponible à l'adresse : https://doc.rero.ch/record/258018/files/TDB_Tieche_Julien.pdf

TRACE, Ciaran B., 2021. Archival infrastructure and the information backlog. *Archival Science* [en ligne]. 21 juillet 2021. pp. 75-93. [Consulté le 29 novembre 2021]. Disponible à l'adresse : https://link.springer.com/content/pdf/10.1007/s10502-021-09368-x.pdf

TREFFEISEN, Jürgen, 2000. Die Transparenz der Archivierung – Entscheidungsdokumentation bei der archivischen Bewertung. BRÜBACH, Nils (éd.), *Der Zugang zu Verwaltungsinformationen – Transparenz als archivische Dienstleistung. Beiträge des 5. Archivwissenschaftlichen Kolloquiums der Archivschule Marburg*. 2000. pp. 177-179.

TÜRCK, Verena, 2014. *Veränderungen von Bewertungsgrundsätzen bei der Übernahme digitaler Unterlagen?* [en ligne]. Transferarbeit Laufbahnprüfung für den Höheren Archivdienst. Marburg : Archivschule Marburg. [Consulté le 26 juin 2022]. Disponible à l'adresse : https://www.landesarchiv-bw.de/sixcms/media.php/120/57173/Transferarbeit_VerenaTuerck_02.pdf

UPWARD, Frank, 1996. Structuring the records continuum part one. Post-custodial principles and properties. *Archives & Manuscripts* [en ligne]. 1 novembre 1996. Vol. 24, n° 2, pp. 268-285. [Consulté le 19 juin 2022]. Disponible à l'adresse : https://publications.archivists.org.au/index.php/asa/article/view/8583

UPWARD, Frank, REED, Barbara, OLIVER, Gillian et EVANS, Joanne, 2018. *Recordkeeping Informatics for a Networked Age*. Clayton, Victoria : Monash Univeristy Publishing. ISBN 978-1-925495-88-1.

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]. Disponible à l'adresse : https://doi.org/10.1108/RMJ-02-2016-0006

VENKATA, Santhilata Kuppili, 2020. *A Benchmarking Tool for AI-for-Selection of Documents for Permanent Preservation* [en ligne]. The National Archives UK. [Consulté le 9 août 2022]. Disponible à l'adresse : https://cdn.nationalarchives.gov.uk/documents/ai-for-selection-national-archives.pdf

VEUVE, Nils, 2021. *Le projet AENeas: Les Archives de l'Etat de Neuchâtel face aux défis de l'archivage numérique*. Travail de Master MAS. Berne et Lausanne : MAS ALIS, Universités de Berne et Lausanne.

VINH-DOYLE, William P., 2017. Appraising email (using digital forensics) : techniques and challenges. *Archives and Manuscripts* [en ligne]. Mars 2017. Vol. 45, n° 1, pp. 18-30. [Consulté le 30 octobre 2021]. Disponible à l'adresse : https://search.informit.org/doi/10.3316/agispt.20171146

WALNE, Peter (éd.), 1988. *Dictionary of archival terminology = Dictionnaire de terminologie archivistique : English and French, with equivalents in Dutch, German, Italian, Russian and Spanish*. 2nd rev. München : K.G. Saur. ICA handbooks series. ISBN 3-598-20279-2.

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]. Disponible à l'adresse : https://www.persee.fr/doc/gazar_0016-5522_1990_num_149_1_3151

WETTMAN, Andrea, 2008. Bewertung per Mausklick : zur Aussonderung und Archivierung elektronischer Akten. *Archive in Bayern* [en ligne]. 2008. Vol. 4, pp. 265-277. [Consulté le 26 juin 2022]. Disponible à l'adresse : https://hds.hebis.de/asmr/Record/HEB405623399

WILDE, Max De et HENGCHEN, Simon, 2016. Semantic Enrichment of a Multilingual Archive with Linked Open Data. *Digit. Human. Quart.* [en ligne]. 1 janvier 2016. Vol. 11, n° 4. [Consulté le 19 juin 2022]. Disponible à l'adresse : https://www.researchgate.net/profile/Simon-Hengchen/publication/295632397_Semantic_Enrichment_of_a_Multilingual_Archive_with_Linked_Open_Data/links/57c92b4608ae9d640483105e/Semantic-Enrichment-of-a-Multilingual-Archive-with-Linked-Open-Data.pdf?origin=publication_detail

WINCATALOG, 2022. Main Features - Disk Catalog Software for Windows - WinCatalog 2021. *WinCatalog 2021* [en ligne]. 2022. [Consulté le 11 août 2022]. Disponible à l'adresse : https://www.wincatalog.com/features.html#virtual-folders

WOLPERT, D.H. et MACREADY, W.G., 1997. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation* [en ligne]. Avril 1997. Vol. 1, n° 1, pp. 67-82. [Consulté le 5 août 2022]. Disponible à l'adresse : https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=585893

WOLPERT, D.H. et MACREADY, W.G., 2005. Coevolutionary free lunches. *IEEE Transactions on Evolutionary Computation* [en ligne]. Décembre 2005. Vol. 9, n° 6, pp. 721-735. [Consulté le 5 août 2022]. Disponible à l'adresse : https://ieeexplore.ieee.org/document/1545946

[Workflow] Flux de travaux. *Grand dictionnaire terminologique* [en ligne]. 2012. [Consulté le 10 août 2022]. Disponible à l'adresse : https://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=8362151

YOAKIM, William, 2022. Les archives sont infiniment humaines : il convient maintenant de les recevoir et de les traiter comme telles. *arbido* [en ligne]. 2022. Vol. 1. [Consulté le 11 juillet 2022]. Disponible à l'adresse : https://arbido.ch/fr/edition-article/2022/archiver-linarchivable/les-archives-sont-infiniment-humaines-il-convient-maintenant-de-les-percevoir-et-de-les-traiter-comme-telles

YOUNG, Tom, HAZARIKA, Devamanyu, PORIA, Soujanya et CAMBRIA, Erik, 2017. Recent Trends in Deep Learning Based Natural Language Processing. [en ligne]. 9 août 2017. [Consulté le 13 janvier 2022]. Disponible à l'adresse : https://arxiv.org/abs/1708.02709v8

ZELLER, Jean-Daniel, 2003. It's a long way to e-archiving.... *arbido* [en ligne]. 2003. Vol. 3, pp. 7-8. [Consulté le 13 août 2022]. Disponible à l'adresse : https://arbido.ch/assets/files/arbido_3.3_001_040.pdf

ZELLER, Jean-Daniel, 2013. Une stratégie et après… Dix ans de développement de l'archivage électronique en Suisse (2002-2012). *Gazette des archives* [en ligne]. 2013. Vol. 229, n° 1, pp. 187-210. [Consulté le 10 juillet 2022]. Disponible à l'adresse : https://www.persee.fr/docAsPDF/gazar_0016-5522_2013_num_229_1_5202.pdf

ZIWES, Franz-Josef, 2020. Bewertung zwischen Fingerspitzengefühl und e-Skills. Strategien zur Bewältigung einer archivischen Kernaufgabe. In : *Aktuelle Fragen der Überlieferungsbildung Vorträge des 79. Südwestdeutschen Archivtags am 16. und 17. Mai 2019 in Ludwigsburg*. Stuttgart : Verlag W. Kohlhammer. 2020. pp. 37-45. ISBN 978-3-17-038171-1.

ZWEIFEL, Josef, 1995. *Die Bewertung im Rahmen der vorarchivischen Schriftgutverwaltung*. 1995. Vol. 9, pp. 19-14.

ZWICKER, Josef, 1995. Archivische Bewertung: ein juristisches Problem? *arbido*. 1995. Vol. 9, pp. 24-27.

ZWICKER, Josef, 2005. Zum Stand der Bewertungsdiskussion in der Schweiz nebst Bemerkungen zu den Aussengrenzen der Überlieferungsbildung. In : BISCHOFF, Frank M. et KRETZSCHMAR, Robert, *Neue Perspektiven archivischer Bewertung: Beiträge zu einem Workshop an der Archivschule Marburg*. Marburg. pp. 101-118.

Annexe 1 : Grille d'entretien – pratiques d'évaluation cantons romands

1. Sur quelle(s) base(s) légale(s) s'appuie vos activités ? l'évaluation y est-elle citée ?
2. Votre service d'archives/sa direction dispose-t-il d'un cadre et d'une approche stratégique de l'évaluation ?
3. Workflow évaluation :
    a. Les services disposent-ils d'un plan de classements ? Sous quelle forme ? Comment sont-ils appliqués ?
    b. Que déclenche une évaluation sur le terrain ?
    c. Qui se déplace ? à quel rythme (approximatif)
    d. Pouvez-vous décrire comment vous procédez lorsque vous vous procédez à une évaluation :
        i. Quelles tâches sont réalisés ?
        ii. Quels intervenants jouent quels rôles ?
    e. Critères d'évaluation
        i. Pouvez-vous les définir dans les grandes lignes ?
        ii. Les critères sont-ils formalisés ? dans un Document/guide ? Mis à jour/tous les combien de temps ?
        iii. Comment la valeur probante, l'exploitabilité et la représentativité (ISO 15489 sur le Records Management) sont-ils jugés ou pris en compte ?
        iv. Échantillonnage : dans quel cas et de quelle manière est-il appliqué ?
        v. Travail au niveau des dossiers ? Les doublons au sein des dossiers sont-ils éliminés ?
    f. Une justification de la décision (versement/élimination est-elle notifiée ?)
    g. Votre service pratique-t-il la réévaluation ?
4. Comment s'effectue le suivi des services ? quel(s) outil(s) sont impliqués ?
5. Les processus sont-ils très différents pour les archives privées ?
6. Qu'en est-il des archives numériques ? comment-procédez-vous ? Quels outils sont utilisés ?
7. Comment jugez-vous le passif d'archives à traiter dans votre canton ?

Annexe 2 : Résultats questionnaire archivistes AEN

Variables	testeur 1	testeur 2	testeur 3	testeur 4	testeur 5	Moy	Moy_inv.	%	Ecart-type
V1. Documentation de la transmission	4	2	1	5	5	3,4	2,6	52	1,82
V.2 Enregistrement des événements liés aux interventions sur le document	2	2	3	5	5	3,4	2,6	52	1,52
V.3 Intégration du document dans un référentiel documentaire	2	1	1	3	1	1,6	4,4	88	0,89
V.4 Complétude des composants du document ou du dossier	2	2	1	3	2	2	4	80	0,71
V.5 Complétudes des métadonnées	2	1	1	5	3	2,4	3,6	72	1,67
V.6 Habilitation du producteur	2	4	1	3	3	2,6	3,4	68	1,14
V.7 Conformité des procédures de traitement de l'activité	2	2	2	5	3	2,8	3,2	64	1,30
V.8 Connaissance du producteur	1	1	1	1	2	1,2	4,8	96	0,45
V.9 Existence d'un identifiant	1	4	3	5	1	2,8	3,2	64	1,79
V.10 Existence d'un intitulé	1	3	1	1	1	1,4	4,6	92	0,89
V.11 Format du nom de fichier	2	4	1	5	5	3,4	2,6	52	1,82
V.12 Indication des dates de création	1	1	1	2	1	1,2	4,8	96	0,45
V.13 Existence d'une signature	2	2	1	5	5	3	3	60	1,87
V.14 Fixité de la chaîne de bits	2	3		5	1	2,75	3,25	65	1,71
V.15 Inscription du document dans une activité officielle	2	1	1	1	1	1,2	4,8	96	0,45
V.16 Exclusivité de l'information	3	1	1	2	3	2	4	80	1,00
V.17 Correspondance des thématiques à un référentiel documentaire	1	1	1	3	1	1,4	4,6	92	0,89
V.18 Pertinence géographique	2	2	2	2	2	2	4	80	0,00
V.19 Pertinence temporelle	2	2	1	2	1	1,6	4,4	88	0,55
V.20 Fréquence des activités	2	3	3	5	3	3,2	2,8	56	1,10
V.21 Validité des chemins d'accès	2	2	3	5	1	2,6	3,4	68	1,52
V.22 Caractéristiques significatives	2	3	3		5	3,25	2,75	55	1,26
V.23 Nature du format du fichier	2	1	2	4	1	2	4	80	1,22
V.24 Application de création	1	1	1	4	1	1,6	4,4	88	1,34
V.25 Environnement logiciel de restitution	2	2	2	4	1	2,2	3,8	76	1,10
V.26 Medium de stockage	2	2	2	5	1	2,4	3,6	72	1,52
V.27 Disponibilité des codes d'accès à l'objet	1	3	1	1	5	2,2	3,8	76	1,79
V.28 Multiplicité des points d'entrées	2	3	3		1	2,25	3,75	75	0,96
V.29 Description du producteur	2	1	1	2	1	1,4	4,6	92	0,55
V.30 Description du contexte de création	2	2	2	2	1	1,8	4,2	84	0,45
V.31 Description du contenu	2	1	3	1	1	1,6	4,4	88	0,89
V.32 Présence des langues officielle	2	2	2	3		2,25	3,75	75	0,50
V.5 Complétudes des métadonnées	2	1	1	5	3	2,4	3,6	72	1,67
V.33 Protection intellectuelle	2	4	3	5	3	3,4	2,6	52	1,14
V.34 Protection des données	1	3	1	5	3	2,6	3,4	68	1,67
V.35 Position du producteur	2	1	1	3		1,75	4,25	85	0,96
V.36 Genre de document	2	1	1	3	1	1,6	4,4	88	0,89
V.37 Relation avec les fonctions du producteur	3	2	3	1	1	2	4	80	1,00
V.38 Diffusion du document	2	3	3	3	4	3	3	60	0,71
V.39 Visibilité du réseau de producteur dans le document	3	1	2	3	1	2	4	80	1,00
V.40 Originalité du contexte	2	2	1	1	1	1,4	4,6	92	0,55
V.41 Qualité esthétique	2	4	1	2	2	2,2	3,8	76	1,10
V.42 Qualité patrimoniale	1	1	1	1	1	1	5	100	0,00

Annexe 3 : Définitions des dimensions d'évaluations et des variables

Comme l'indique la figure 3 ci-dessus, les dimensions d'évaluation amènent à l'émergence de variables, véritables éléments de mesures qui permettent de quantifier leurs valeurs associées, de vérifier leurs mesures et, enfin, de les reproduire. L'émergence de ces variables a été rendue possible grâce à la conception d'un cadre conceptuel qui décompose les trois dimensions principales en éléments plus spécifiques et ainsi plus facilement appréhendables. De la sorte, les DEVs ont été décomposés à l'aide de trois niveaux : DEV1, DEV2, et DEV3.

DEV1. Valeur probante

Définition : caractéristique ou propriété, pour un document, démontrant sa capacité à prouver une opération (Organisation internationale de normalisation 2016, p. 2 ; Université Laval 2014).

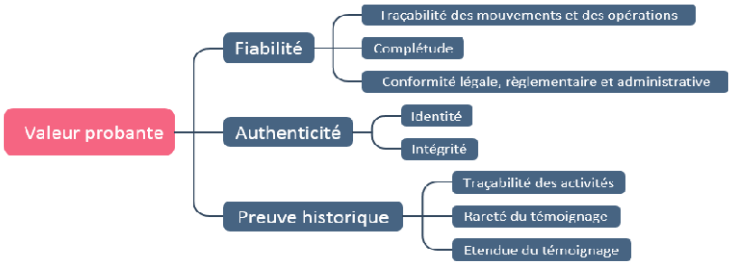


Figure 4 : DEV1. Valeur probante

DEV2	*DEV3*	*Variable*
Fiabilité « Un document d'activité fiable est un document dont le contenu peut être considéré comme la représentation complète et exacte des opérations, activités ou faits qu'il atteste, sur lequel on peut s'appuyer lors d'opérations ou d'activités ultérieures » (Organisation internationale de normalisation 2016, p.5)		
	Traçabilité des mouvements et des opérations « Fait de créer, d'enregistrer et de préserver les données relatives aux mouvements et à l'utilisation des documents d'activité » (Organisation internationale de normalisation 2011, p.11)	
		V1. Documentation de la transmission Information relative au transfert de l'objet de l'entité organisationnelle à l'entité d'archives (Makhlouf Shabou, Mellifluo et Rey 2013, p. 20) **V2. Enregistrement des événements liés aux interventions sur le document** Cette variable évalue la façon dont ont été consignés les événements affectant un document tout au long de son existence (Organisation internationale de normalisation 2016, p. 6) **V3. Intégration du document dans un référentiel documentaire** Cette variable évalue l'utilisation de référentiels documentaires permettant de relier les documents à leur contexte de création (Organisation internationale de normalisation 2016, p. 2)

Complétude
Un document possédant tous les éléments requis par son producteur ou ses contraintes légales. Sous-entend la capacité à renseigner de façon autonome sur une affaire (InterPARES [s. d.] ; Makhlouf Shabou 2010, p. 158)

V4. Complétude des composants du document
Cette variable évalue la capacité pour un document à renseigner de façon autonome sur une affaire spécifique (Makhlouf Shabou 2010, p. 158)

V.5. Complétudes des métadonnées
Cette variable évalue la capacité pour un document à donner l'ensemble des informations requises au travers de ses métadonnées

Conformité légale, réglementaire et administrative
La conformité légale, réglementaire et administrative repose sur les procédures de création et sur le producteur (Makhlouf 2010, p. 158). « Soit sur la personne physique ou morale qui crée, reçoit, ou assemble des documents d'activité dans le cadre de ses missions, fonctions ou activités, et sur les procédures qui amènent la création de documents d'activité » (Tièche 2015, p. 53)

V6. Habilitation du producteur
Cette variable évalue la capacité pour un document à donner des informations relatives à son producteur (Makhlouf Shabou 2010, p. 118 ; Tièche, p. 53)

V7. Conformité des procédures de traitement de l'activité
Cette variable évalue la capacité pour un document à donner des informations relatives aux procédures à l'origine de sa création (Makhlouf Shabou, 2010, p. 118 ; Tièche 2015, p. 53)

Authenticité
« Un document d'activité authentique est un document dont on peut prouver qu'il est bien ce qu'il prétend être, qu'il a été créé ou envoyé par l'acteur qui prétend l'avoir créé ou envoyé, et qu'il a été créé ou envoyé au moment prétendu » (Organisation internationale de normalisation 2016, p.5)

Identité
Ensemble des caractéristiques qui permettent d'identifier un document ou document d'archives de façon unique, et par là même de la distinguer des autres (InterPARES [s. d.])

V8. Connaissance du producteur
Cette variable évalue la présence d'une information identifiant son producteur (Makhlouf Shabou, Mellifluo et Rey 2013, p. 9)

V9. Existence d'un identifiant
Cette variable évalue la présence d'un ou plusieurs identifiants permettant d'identifier le document de façon unique (Organisation internationale de normalisation 2012b, p. 33)

V10. Existence d'un intitulé
Cette variable évalue la présence d'un titre rattaché au document (Makhlouf Shabou, Mellifluo et Rey 2013, p. 6)

V11. Format du nom de fichier
Cette variable évalue l'existence de règles de nommage pour les noms de fichiers et sa mise en conformité (Makhlouf Shabou, Mellifluo et Rey 2013, p. 7)

V12. Indication des dates de création
Cette variable évalue la présence d'une information relative à la date de création et/ou d'enregistrement du document (Makhlouf Shabou, Mellifluo et Rey 2013, p. 11)

V13. Existence d'une signature
Cette variable évalue la présence d'une signature permettant d'authentifier le producteur (Tièche 2015, p. 48)

Intégrité
« L'intégrité d'un document d'activité renvoie au caractère complet et non altéré de son état » (Organisation internationale de normalisation 2016, p.5)

V14. Fixité de la chaîne de bits
Cette variable évalue la présence d'une empreinte numérique permettant de s'assurer que le document n'ait pas été modifié de façon non documentée (Organisation internationale de normalisation 2012b, p. 74)

Preuve historique
La preuve historique correspond à l'importance du document dans sa fonction de témoin d'un contexte historique de la société ou de l'organisation à l'origine de sa création (Makhlouf Shabou, Mellifluo et Rey 2013, p. 25)

Traçabilité des activités
Selon la loi sur l'archivage, la traçabilité des activités des autorités cantonales est un des buts de l'archivage des documents. La capacité à répondre à ce but impacte la valeur archivistique des documents (LArch, art. 2)

V15. Inscription du document dans une activité officielle
Cette variable évalue la présence d'une information permettant de relier le document à une activité officielle de l'organisation

Rareté du témoignage
Correspond à la rareté des informations contenus dans le document et de l'existence des celles-là dans d'autres sources (Makhlouf Shabou, Mellifluo et Rey 2013, p. 25)

V16. Exclusivité de l'information
Cette variable évalue la rareté de l'information contenu dans un document en regard de l'existence d'autres sources d'information sur la même affaire (Makhlouf Shabou, Mellifluo et Rey 2013, p. 25)

Etendue du témoignage
« Eléments couverts par le témoignage à la fois sur le plan thématique, temporel et local » (Makhlouf Shabou, Mellifluo et Rey 2013, p. 26)

V17. Correspondance des thématiques à un référentiel documentaire
Cette variable permet d'attribuer une valeur supplémentaire à certains documents reliés à des thématiques particulières (Makhlouf Shabou, Mellifluo et Rey 2013, p. 26)

V18. Pertinence géographique
Cette variable évalue la relation entre le document et une aire géographique donnée (Makhlouf Shabou, Mellifluo et Rey 2013, pp. 26-27)

V19. Pertinence temporelle
Cette variable évalue la relation entre le document et une période chronologique donnée (Makhlouf Shabou, Mellifluo et Rey 2013, p. 27)

V20. Fréquence des activités
Cette variable évalue la répétition de l'activité à laquelle est rattaché le document (Makhlouf Shabou, Mellifluo et Rey 2013, pp. 27-28)

DEV1. Exploitabilité

Définition : « un document d'activité exploitable est un document qui peut être localisé, récupéré, communiqué et interprété dans une période de temps jugée raisonnable par les parties prenantes » (Organisation internationale de normalisation 2016, p.5)

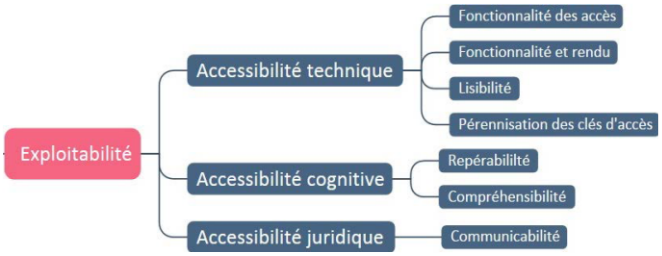


Figure 5 : DEV1. Exploitabilité

DEV2	*DEV3*	*Variable*
Accessibilité technique Facilité d'accéder au contenu d'information, soit au contenu du document archivé (Tièche 2015, p. 53 ; Makhlouf Shabou 2010, p. 122 ; Organisation internationale de normalisation 2012b, p. 21)		
	Fonctionnalité des accès « Capacité technique à repérer un document et à y accéder » (Makhlouf Shabou, Mellifluo et Rey 2013, p. 29)	
		V21. Validité des chemins d'accès Cette variable évalue la capacité technique à accéder ou à repérer un document (Makhlouf Shabou 2010, p. 122 ; Makhlouf Shabou, Mellifluo et Rey, p. 29)
	Fonctionnalité et rendu Caractéristiques d'un document qui doivent être conservées sur le long terme afin de garantir un accès, une utilisation et une signification permanente, ainsi que la capacité pour un document d'être accepté comme preuve de ce qu'il prétend être (Wilson 2008, p. 15)	
		V22. Caractéristiques significatives Cette variable évalue la possibilité de conserver les caractéristiques définies comme fondamentales à travers le temps (Tièche 2015, pp. 56-57)
	Lisibilité « Possibilité de lire les documents grâce à la qualité de leur état physique ainsi que la disponibilité du matériel approprié à leur lecture » (Makhlouf Shabou 2010, p. 122)	
		V23. Nature du format du fichier Cette variable évalue l'existence d'une information permettant de renseigner le format de données du document (Banat-Berger, Duplouy, Huc 2009, p. 47)

V24. Application de création
Cette variable évalue l'existence d'une information permettant de renseigner l'application à l'origine de la création du document (PREMIS Editorial Committee 2012, p. 63)
V25. Environnement logiciel de restitution
Cette variable évalue l'existence d'une information permettant de renseigner quel(s) logiciel(s) permet(tent) de restituer l'information du document (Roussel 2012-2014, pp. 62-63)
V26. Medium de stockage
Cette variable évalue l'existence d'une information permettant de renseigner sur le support de stockage sur lequel se trouve le document. La connaissance de cette information permet d'anticiper les migrations de supports (Banat-Berger, Duplouy, Huc 2009, pp. 28, 82, 120-121 ; PREMIS Editorial Committee 2012, p. 79)

Pérennisation des clés d'accès
« Conservation à long terme des clés permettant d'accéder au contenu des documents » (Makhlouf Shabou, Mellifluo et Rey 2013, p. 32)

V27. Disponibilité des codes d'accès à l'objet
Cette variable évalue l'existence de clés d'accès ou de systèmes de hachage requis pour accéder au contenu du document (Makhlouf Shabou, Mellifluo et Rey 2013, p. 32)

Accessibilité cognitive
Accès aisé au contenu et au contexte de création des archives (Boles et Young, 1991, cité dans Makhlouf Shabou 2010, p. 120).

Repérabilité
Capacité, pour un document, d'être identifiable et recherchable (Organisation internationale de normalisation 2016, p.1)

V28. Multiplicité des points d'entrées
Cette variable évalue l'existence d'informations relatives aux moyens cognitifs d'accéder au document (Makhlouf Shabou, Mellifluo et Rey 2013, pp. 33-34)

Compréhensibilité
Capacité et aisance à appréhender le contenu d'un document (Makhlouf Shabou 2010, p. 121)

V29. Description du producteur
Cette variable évalue l'existence d'informations relatives au producteur (Makhlouf Shabou, Mellifluo et Rey 2013, p. 35)
V30. Description du contexte de création
Cette variable évalue l'existence d'informations relatives au contexte de création du document. Il s'agit pour la plupart des cas d'informations relatives à la législation à l'origine de la fonction responsable de la création du document (Makhlouf Shabou, Mellifluo et Rey 2013, p. 35-36 ; Conseil international des archives 2007, p. 18)
V31. Description du contenu
Cette variable évalue la présence d'information permettant de comprendre le document dans son contexte (Makhlouf Shabou, Mellifluo et Rey 2013 p. 35). La norme ISO 14721 demande à ce que l'information à conserver soit compréhensible de façon autonome par sa communauté d'utilisateurs (Organisation internationale de normalisation 2012b, p.39)

V32. Présence des langues officielles
Cette variable évalue la présence ou l'absence de(s) langue(s) officielle(s) de l'organisation (Makhlouf Shabou, Mellifluo et Rey 2013, p. 37)
V5. Complétudes des métadonnées
Cette variable évalue la capacité pour un document à donner l'ensemble des informations requises au travers de ses métadonnées

Accessibilité juridique
« Communicabilité des [documents] qui s'appuient sur l'autorisation juridique de [les consulter] » (Makhlouf Shabou 2010, p. 121)

Communicabilité
Absence de contraintes légales ou réglementaires empêchant la consultation et l'utilisation du document (Makhlouf Shabou 2010, pp. 120-121 ; Tièche, TB 2015, p. 58)

V33. Protection intellectuelle
Cette variable évalue la présence d'informations relatives à l'existence de dispositifs, tel le droit d'auteur, restreignant la diffusion et l'exploitation du document (Tièche 2015, p. 59)
V34. Protection des données
Cette variable évalue la présence d'informations relatives à l'existence de délai de protection restreignant la consultation, la diffusion et l'exploitation du document (Tièche 2015, p. 59)

DEV1. Représentativité

Définition : « Capacité des archives à permettre un témoignage significatif, riche et exhaustif des différents éléments du contexte organisationnel de leur création » (Makhlouf Shabou 2010, p. 123)

Figure 6 : DEV1. Représentativité

DEV2	*DEV3*	*Variable*
Contexte organisationnel		
Capacité des documents à exprimer l'importance de l'organisation et de ses unités, et les liens qui le caractérisent (Makhlouf Shabou 2010, p. 125 ; Menne-Haritz 1994 ; Cook, 2005)		
	Significativité du producteur	
	« Représentativité du document reflétée à travers le contexte et l'importance de son producteur » (Makhlouf Shabou, Mellifluo et Rey 2013, p. 40)	
		V35. Position du producteur
		Cette variable évalue la position du producteur, qu'elle soit au sein de l'organisation ou en dehors (Makhlouf Shabou, Mellifluo et Rey 2013, pp. 40-41)

Significativité du document
« Représentativité du document reflétée à travers sa propre nature, ses usages et ses liens avec le contexte de production » (Makhlouf Shabou, Mellifluo et Rey 2013, p. 41)

V36. Genre de document
Cette variable évalue la fonction du document. Un document stratégique peut représenter une valeur différente d'un document informatif (Makhlouf Shabou, Mellifluo et Rey 2013, pp. 41-42)

V37. Relation avec les fonctions du producteur
Cette variable évalue si le document est issu d'une activité spécifique de l'organisation, ou s'il s'agit d'une activité de gestion présente dans l'ensemble des d'organisation telle la gestion des ressources humaines (Makhlouf Shabou, Mellifluo et Rey 2013, p. 42)

V38. Diffusion du document
Cette variable évalue l'existence d'informations relatives à la circulation du document, et tout particulièrement s'il s'agit d'une diffusion interne ou publique (Makhlouf Shabou, Mellifluo et Rey 2013, p. 41)

V39. Visibilité du réseau de producteur dans le document
Cette variable évalue le témoignage du fonctionnement de l'organisation productrice du document, notamment dans sa dimension relationnelle avec d'autres acteurs (Makhlouf Shabou, Mellifluo et Rey 2013, p. 41)

Contexte socio-culturel
Capacité du document à témoigner du contexte socio-culturel de l'organisation productrice (Makhlouf Shabou, Mellifluo et Rey 2013, p. 41)

Rareté contextuelle
Rareté en lien avec la date de création, le producteur, ou encore les raisons de création du document (Makhlouf Shabou, Mellifluo et Rey 2013, p. 43)

V40. Originalité du contexte
Cette variable évalue l'intérêt que peut présenter le contexte de création du document (Makhlouf Shabou, Mellifluo et Rey 2013, p. 43)

Valeur esthétique
Valeur du document en lien avec son caractère de beauté ou artistique (Makhlouf Shabou, Mellifluo et Rey 2013, p. 44)

V41. Qualité esthétique
Cette variable évalue la qualité esthétique du document représentée par le document et son contenu (Makhlouf Shabou, Mellifluo et Rey 2013, p. 44)

Valeur patrimoniale
Valeur du document relative à sa qualité patrimoniale

V42. Qualité patrimoniale
Cette variable évalue la qualité patrimoniale du document représentée par le document et son contenu

Les sections précédentes ont exposé la synthèse de notre réflexion sur les variables qui pourraient être examinées pour l'identification des métriques et des mesures concrètes à retenir dans l'opérationnalisation de l'outil 4. Ces 42 variables seront étudiées et analysées en vue d'en sélectionner les plus automatisables.

BIBLIOGRAPHIE[1]

ABUZAWAYDA, Yousef I., YUSOF, Zawiyah M., AB AZIZ, Mohd Juzaiddin, 2012. Significance of automated records retention schedule among universities in Malaysia. In : *Seventh International Conference on Digital Information Management (ICDIM), Macao, Chine, 22-24 août 2012* [en ligne]. Institute of Electrical and Electronics Engineers (IEEE), 26 novembre 2012. [Consulté le 26 janvier 2018]. Disponible à l'adresse : 10.1109/ICDIM.2012.6360143

OFFICE DES ARCHIVES DE L'ÉTAT DE NEUCHÂTEL, 2016. *Projet AENeas : Description du projet AENeas basé sur le concept*. Neuchâtel : Archives de l'État, novembre 2016.

BANAT-BERGER, Françoise, DUPLOUY, Laurent, HUC, Claude, 2009. *L'archivage numérique à long terme : les débuts de la maturité ?* Paris : Direction des Archives de France, 2009. Manuels et guides pratiques. ISBN 978-2-11-006942-9

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

CONSEIL INTERNATIONAL DES ARCHIVES, [s. d.]. *Multilingual Archival Terminology* [en ligne]. [Consulté le 30 janvier 2018]. Disponible à l'adresse : http://www.ciscra.org/mat/mat

COOK, Terry, 2005. Macroappraisal in Theory and Practice: Origins, Characteristics, and Implementation in Canada, 1950-2000. *Archival Sciences* [en ligne]. Décembre 2005. Vol. 5, pp. 101-161. [Consulté le 30 janvier 2018]. Disponible à l'adresse : https://link.springer.com/article/10.1007/s10502-005-9010-2

COUTURE, Carol, 1999. Évaluation. *Les fonctions de l'archivistique contemporaine*. Québec : Presse de l'Université du Québec, pp. 103-143

CRAIG, Barbara, 2004. *Archival Appraisal. Theory and Practice*. Munich : K.G.Saur. ISBN 3-598-11538-5

DOOM, Vincent, 2006. L'évaluation scientifique des archives : principes et stratégies : du melon au diamant. *La Gazette des Archives* [en ligne]. 2006, vol. 2, n° 202, pp. 5-43. [Consulté le 25 janvier 2018]. Disponible à l'adresse : http://www.persee.fr/doc/gazar_0016-5522_2006_num_202_2_3815

FORTIN, Marie-Fabienne et GAGNON, Johanne, 2016. *Fondements et étapes du processus de recherche : Méthodes quantitatives et qualitatives*. 3e éd. Montréal : Chenelière Éducation.

INTERPARES TRUST, [s. d.]. *InterPARES Project: Terminology Database* [en ligne]. [Consulté le 30 janvier 2018]. Disponible à l'adresse : http://arstweb.clayton.edu/interlex/index.php

MAKHLOUF SHABOU, Basma, 2009. La contribution des principes de l'évaluation archivistique aux qualités des archives définitives = The Contribution of the Appraisal Principles to the Qualities of Historical Archival. Encontros Bibli [en ligne]. [Consulté le 17 février 2017]. Disponible à l'adresse : http://www.redalyc.org/pdf/147/14712771008.pdf

MAKHLOUF SHABOU, Basma, 2010. *Étude sur la définition et la mesure des qualités des archives définitives issues d'une évaluation* [en ligne]. Montréal : École de bibliothéconomie et des sciences de l'information. Thèse. [Consulté le 9 janvier 2018]. Disponible à l'adresse : https://papyrus.bib.umontreal.ca/xmlui/handle/1866/4955

[1] Nous tenons à remercier M. Aurèle Nicolet, assistant HES à la HEG, pour sa collaboration précieuse à l'élaboration de la bibliographie.

MAKHLOUF SHABOU, Basma, 2015a. Fonctions d'évaluation des archives : bilan sommaire des développements, des enjeux actuels et des défis futurs. IN : Sous la dir. de Louise Gagnon-Arguin et Marcel Lajeunesse. *Panorama de l'archivistique contemporaine. Évolution de la discipline et de la profession. Mélanges offerts à Carol Couture*. Québec : Presses Universitaires du Québec, 2015. pp. 195-214, ISBN 978-2-7605-4337-9

MAKHLOUF SHABOU, Basma, 2015b. Digital Diplomatics and Measurement of Electronic Public Data Qualities: What lessons should be learned?. *Records Management Journal*, 2015. Vol. 25, n° 1. [Consulté le 30 janvier 2018]. Disponible à l'adresse : http://www.emeraldinsight.com/doi/full/10.1108/RMJ-01-2015-0006

MAKHLOUF SHABOU, Basma, MELLIFLUO, Laure et REY, Raphaël (colabor.), 2013. *QADEPs : Définition et mesure des qualités des archives et documents électroniques publics*. Rapport scientifique. Genève : Haute école de gestion, juillet 2013.

MARSHALL, Jennifer Alycen, 2007. *Accounting for Disposition: A Comparative Case Study of Appraisal Documentation at the National Archives and Records Administration in the United States, Library and Archives Canada, and the National Archives of Australia* [en ligne]. University of Pittsburg, 2 février 2007. Thèse de doctorat. [Consulté le 31 janvier 2018]. Disponible à l'adresse : http://d-scholarship.pitt.edu/6251/

MELLIFLUO, Laure, 2008. *Évaluation des archives : en théorie et en pratique aux archives communales de la ville de Carouge* [en ligne]. Genève : Haute de Gestion. [Consulté le 17 février 2017]. Disponible à l'adresse : http://doc.rero.ch/record/11281?ln=fr

Menne-Haritz, Angelika, 1994. Appraisal or Selection. Can Content Oriented Appraisal be Harmonised with the Principle of Provenance? *The Principle of Provenance: Report from the First Stockholm Conference on Archival Theory and the Principle of Provenance, 2-3 September 1993*. Stockholm : Swedish National Archives, pp. 103-131.

THE NATIONAL ARCHIVES, 2013. *Best practice guide to appraising and selecting records for The National Archives* [en ligne]. 15 mars 2013. [Consulté le 17 février 2017]. Disponible à l'adresse : http://www.nationalarchives.gov.uk/documents/information-management/best-practice-guide-appraising-and-selecting.pdf

PREMIS EDITORIAL COMMITTEE, 2012. *PREMIS Data Dictionary for Preservation Metadata* [en ligne]. [s.l.] : PREMIS Editorial Committee, juillet 2012. Version 2.2. [Consulté le 29 janvier 2018]. Disponible à l'adresse : www.loc.gov/standards/premis/v2/premis-2-2.pdf

ROUSSEL, Stéphanie, 2012-2014. Le champ normatif de l'archivage électronique. In: Normalisation et gestion des documents d'activité (records management) : enjeux et nouvelles pratiques pour notre profession. *La Gazette des Archives* [en ligne]. 2012. N° 228, pp. 59-76. ISSN 0016-5522 [Consulté le 29 janvier 2018]. Disponible à l'adresse : http://www.persee.fr/doc/gazar_0016-5522_2012_num_228_4_4984

SHEPHERD, Elizabeth, YEO, Geoffrey, 2003. Managing appraisal, retention and disposition. In : Managing records: a handbook of principles and practice. London : Facet. pp. 146-172.

TIÈCHE, Julien, 2015. La mesure des dimensions de la qualité des archives électroniques : apport des textes normatifs en matière d'archivage électronique à long terme [en ligne]. Genève : Haute école de gestion. Travail de bachelor. [Consulté le 29 janvier 2018]. Disponible à l'adresse : http://doc.rero.ch/record/258018?ln=fr

UNIVERSITÉ LAVAL, 2014. Valeur. *Dico-wiki archivistique* [en ligne]. 28 avril 2014. [Consulté le 22 janvier 2018]. Disponible à l'adresse : https://www.wiki.archivesnumeriques.hst.ulaval.ca/index.php?title=Valeur&oldid=426

WILSON, Andrew, 2008. Significant Properties of Digital Objects. In : *What to preserve? Significant Properties of Digital Objects. Londres, British Library Conference Centre, 7 avril 2008* [en ligne]. Digital Preservation Coalition.[Consulté le 29 janvier 2018]. Disponible à l'adresse : http://www.dpconline.org/docs/miscellaneous/events/142-presentation-wilson/file

Lois et règlements

Convention intercantonale relative à la protection des données et à la transparence dans les cantons du Jura et de Neuchâtel (CPDT-JUNE ; 150.30). *Recueil systématique de la législation neuchâteloise (RSN)* [en ligne]. 9 mai 2012. Mise à jour le 1er janvier 2013. [Consulté le 22 janvier 2018]. Disponible à l'adresse : http://rsn.ne.ch/DATA/program/books/20134/pdf/15030.pdf

Ordonnance du DFF du 11 décembre 2009 concernant les données et informations électroniques (OeIDI ; 641.201.511). *Les autorités fédérales de la Confédération suisse* [en ligne]. 11 décembre 2009. Version du 1er janvier 2010. [Consulté le 30 janvier 2018]. Disponible à l'adresse : https://www.admin.ch/opc/fr/classified-compilation/20092054/201001010000/641.201.511.pdf

Loi sur l'archivage (LArch ; 442.20). *Recueil systématique de la législation neuchâteloise (RSN)* [en ligne]. 22 février 2011. Mise à jour le 1er janvier 2012. [Consulté le 22 janvier 2018]. Disponible à l'adresse : http://rsn.ne.ch/DATA/program/books/20132/pdf/44220.pdf

Règlement d'exécution de la loi sur l'archivage (442.23). *Recueil systématique de la législation neuchâteloise* (RSN) [en ligne]. 29 avril 2013. Mis à jour le 1er août 2013. [Consulté le 22 janvier 2018]. Disponible à l'adresse : http://rsn.ne.ch/DATA/program/books/20154/pdf/44223.pdf

Normes

CONSEIL INTERNATIONAL DES ARCHIVES, 2007. ISDF : Norme internationale pour la description des fonctions [en ligne]. Paris : CIA. Disponible à l'adresse : https://www.ica.org/sites/default/files/CBPS_2007_Guidelines_ISDF_First-edition_FR.pdf

ORGANISATION INTERNATIONALE DE NORMALISATION, 2006. Information et documentation — Processus de gestion des enregistrements — Métadonnées pour les enregistrements — Partie 1: Principes. Genève : ISO, 15 janvier 2006. ISO 23081-1

ORGANISATION INTERNATIONALE DE NORMALISATION, 2008. Information et documentation — Analyse des processus pour la gestion des informations et documents d'activité. Genève : ISO, 15 juin 2008. ISO/TR 26122

ORGANISATION INTERNATIONALE DE NORMALISATION, 2009. Information and documentation — Managing metadata for records — Part 2: Conceptual and implementation issues. Genève : ISO, 1er juillet 2009. ISO 23081-2

ORGANISATION INTERNATIONALE DE NORMALISATION, 2011a. Information et documentation - Systèmes de gestion des documents d'activité - Principes essentiels et vocabulaire. Genève : ISO, 15 décembre 2011. ISO 30300

ORGANISATION INTERNATIONALE DE NORMALISATION, 2011b. *Information et documentation — Systèmes de gestion des documents d'activité — Exigences*. Genève : ISO, 15 novembre 2011. ISO 30301

ORGANISATION INTERNATIONALE DE NORMALISATION, 2011c. Technologies de l'information — Techniques de sécurité — Gestion des risques liés à la sécurité de l'information. 2e éd. Genève : ISO, 1er juin 2011. ISO 27005

ORGANISATION INTERNATIONALE DE NORMALISATION, 2012a. *Electronic archiving — Part 1: Specifications concerning the design and the operation of an information system for electronic information preservation*. Genève : ISO, 1er février 2012. ISO 14641-1

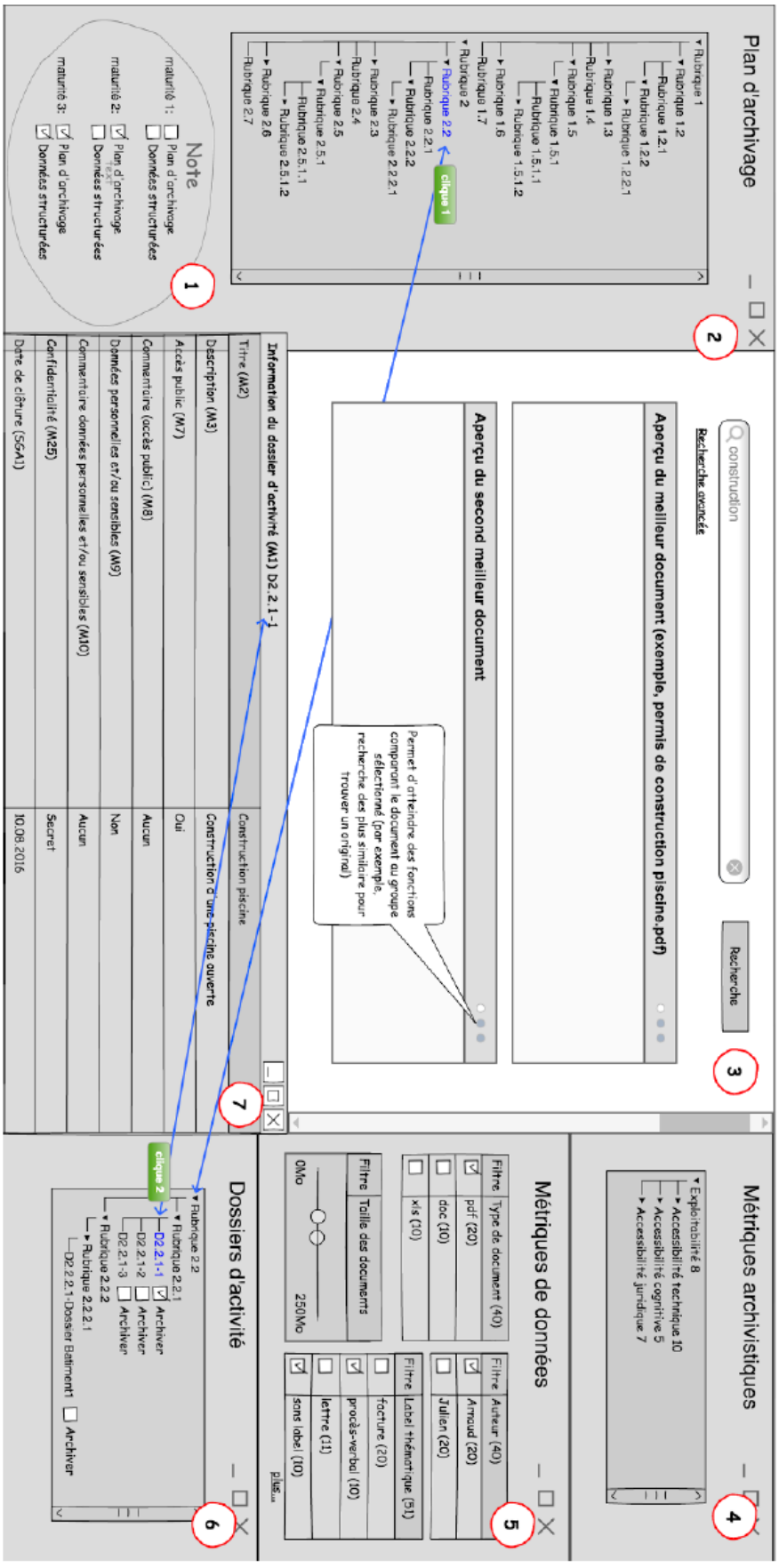
ORGANISATION INTERNATIONALE DE NORMALISATION, 2012b. *Space data and information transfer systems — Open archival information system (OAIS) — Reference model*. 2e éd. Genève : ISO, 1er septembre 2012. ISO 14721

ORGANISATION INTERNATIONALE DE NORMALISATION, 2014. *Information et documentation — Evaluation du risque pour les processus et systèmes d'enregistrement*. Genève : ISO, 15 mars 2014. ISO/TR 18128

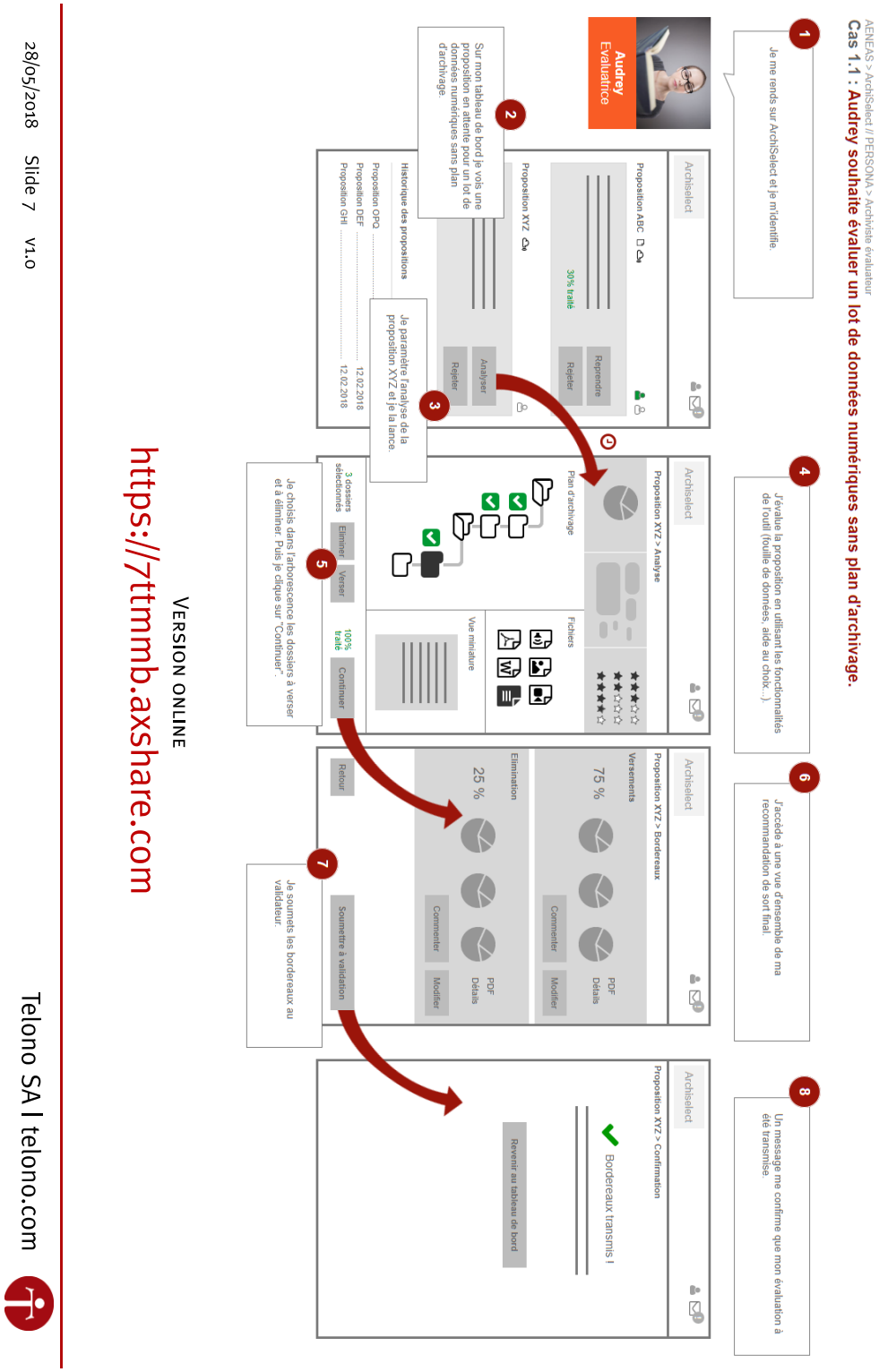
ORGANISATION INTERNATIONALE DE NORMALISATION, 2016. *Information et documentation - Gestion des documents d'activité - Partie 1 : Concepts et principes*. 2e éd. Genève : ISO, 15 avril 2016. ISO 15489-1

ORGANISATION INTERNATIONALE DE NORMALISATION, 2017*. *Information and documentation - Appraisal for managing records*. ISO/AWI TR 21946 [*Note: document interne en préparation].

Annexe 4 : Maquette d'interface pour la fouille de données



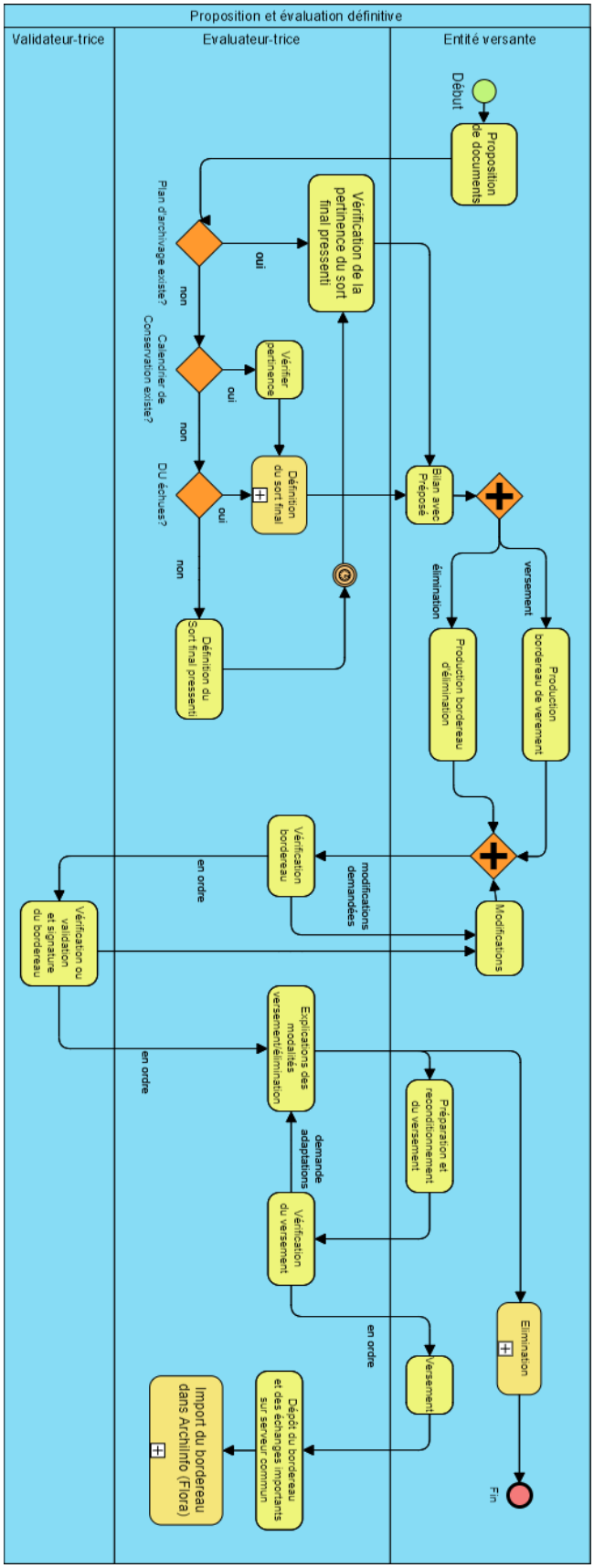
Annexe 5 : Scénario *persona* évaluateur



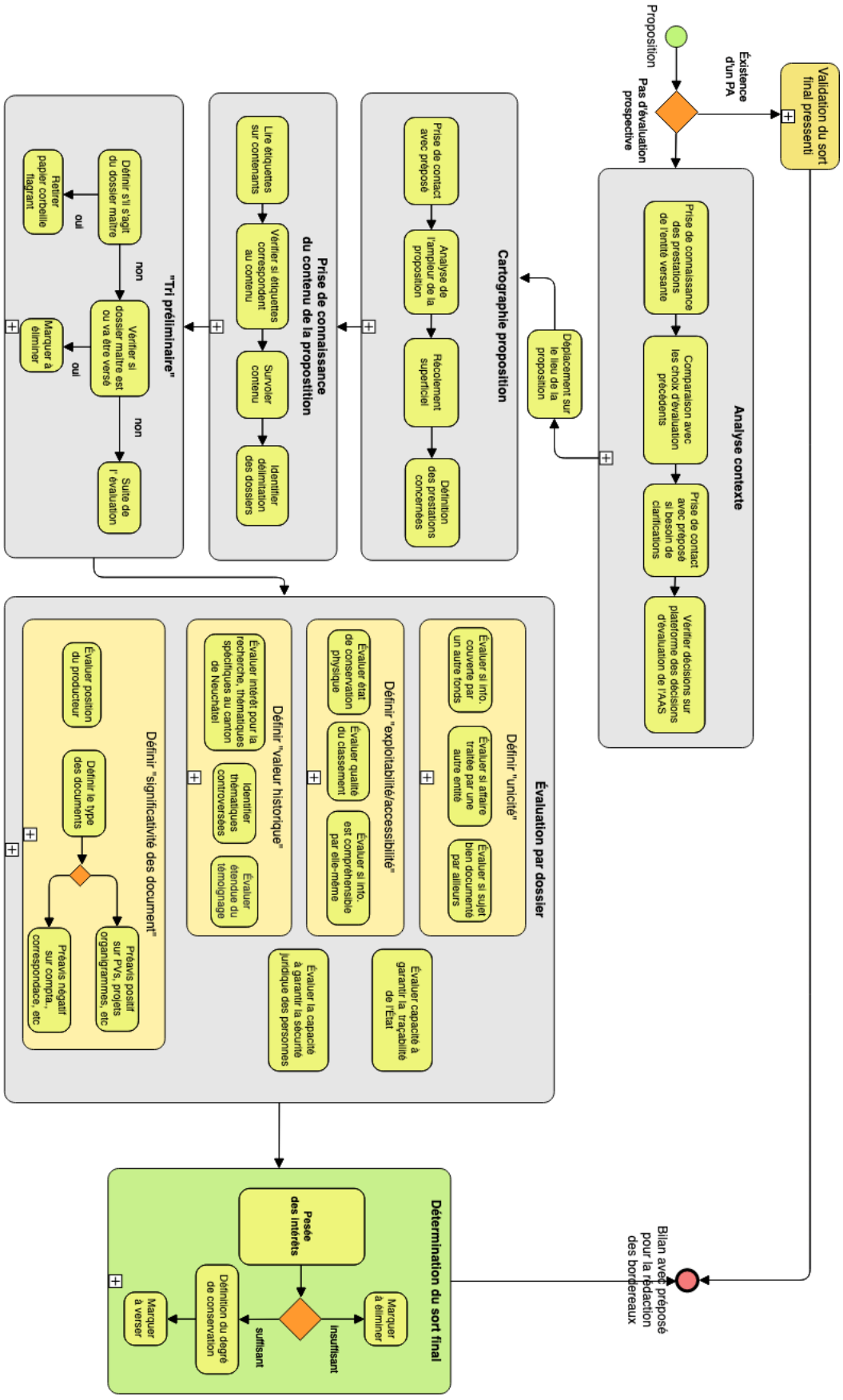
VERSION ONLINE

https://7ttmmb.axshare.com

Annexe 6 : *Workflow* proposition de documents



Annexe 7 : *Workflow* d'évaluation archivistique – analogique



Annexe 8 : Tableau des fonctionnalités proposées

ID	Fonctionnalité	Objectif(s)	*Input*	*Output*	équivalent mandat HEG
f1	Calcul de somme de contrôle	Création d'une empreinte de ce qui a été versé, permettant de vérifier l'intégrité des données aux différentes étapes d'évaluation	Lot de données proposées	Empreinte pour chaque fichier	
f2	Bloqueur d'écriture	Empêcher les fichiers d'être modifiés ou altérés lors du processus d'évaluation	Lot de données proposées	Données bloquées	
f3	Copie de sauvegarde	Création d'une copie de sauvegarde dans le cas de figure que la "zone contrôlée" subit des dommages	Lot de données proposées	Copie de sauvegarde	
f4	Reconnaissance optique de caractères (OCR)	Reconnaissance des documents scannés	Document scanné	Texte reconnu	ana_OCR
f5	Importation données sous forme de SyFi	Importer des lots de données sous forme de fichiers	Lot de données proposées	Système de fichiers accessible par ArchiSelect	preana_import_données
f6	Importation plan d'archivage	Importer le plan d'archivage, ajout des métadonnées du PA dans l'index	Plan d'archivage	Métadonnées du PA dans l'index	preana_import_plan
f7	Importation d'un export SD	Importer un export d'un SD dans l'outil 4 pour ajouter les métadonnées dans l'index	Fichier(s) ad hoc XML, JSON ou YML qui contient les métadonnées	Métadonnées de l'export d'un SD dans l'index pour chaque fichier	preana_import_ged
f8	Définition du niveau de maturité	Définir le niveau de maturité de la gestion documentaire afin de définir le niveau d'analyse	Informations sur outils d'évaluation prospective	Indication niveau 1, 2 ou 3	
f9	Indexation automatique	Indexer un système de fichier dans un moteur de recherche de données structurées et de texte pour offrir des paradigmes de recherche, d'analyses et d'explorations avancées	Un système de fichiers et/ou des métadonnées extraites	Métadonnées et texte dans un index de moteur de recherche	preana_indexation
f10	Visualisation arborescence	Appréhender structure d'un lot de données	Chemins des dossiers	Cartographie de la structure	
f11	Analyse volumétrie	Appréhender volumétrie d'un lot de données	Lot de données proposées	Indicateurs clés de volumétrie	
f12	Inventaire formats	Appréhender formats d'un lot de données	Lot de données proposées	Liste des formats	
f13	Déduplication	Supprimer doublons	Lot de données proposées	Clé de recherche identique	outil_déduplication
f14	Identification des dossiers vides	Identifier dossiers vides	Lot de données proposées	Dossiers vides marqués	
f15	Identification des fichiers temporaires	Identifier fichiers temporaires	Lot de données proposées	Fichiers temporaires marqués	
f16	Sélection par bloc basée sur la cartographie	Pouvoir sélectionner dans la visualisation d'arborescence des blocs de données pour analyse ou indexation	Cartographie	Bloc sélectionné	
f17	Recherche moteur	Chercher des documents et des métadonnées à partir de combinaisons de mots et/ou de métadonnées	Requêtes	Liste des meilleures documents pour les requêtes	rech_moteur
f18	Recherche par facette	Filtrer les recherches avec certaines métadonnées	Requête filtre	Liste des meilleures documents pour les requêtes	rech_filtre
f19	Prévisualisation	Pouvoir prévisualiser un fichier sélectionné	Fichier sélectionnée	Prévisualisation du fichier	
f20	Indexation dynamique	Durant l'exploration pouvoir continuer l'indexation de fichiers ou blocs de fichiers	Fichier	étiquette attribuée	
f21	Reconnaissance d'entités nommées (NER)	Reconnaître les entités nommées dans le texte comme noms, lieux ou institutions	Texte	Liste d'entités nommées (éventuellement position dans le texte)	ana_contenu_rec
f22	Classification typologique (automatique)	Classifier les documents selon une catégorie	Texte et éventuellement autres métadonnées	Liste de classe avec un indice de confiance	ana_contenu_class_texte
f23	*Topic modeling*	Appréhender rapidement les thématiques d'un corpus	Texte	Topic model	
f24	Métriques archivistiques	Obtenir un "score" des DEVs de manière automatique	Métadonnées	Indicateur de confiance	
f25	Visualisation document sur le temps	Voir le nombre de documents dans le temps	Documents datés et périodes de temps	Nombre de documents par période de temps	stat_temps
f26	Détecter cadre de classement	Détecter si un cadre de classement est en place en analysant le chemin du fichier	Système de fichier	Nom et id des dossiers d'activité	ana_fichier_liste_id_titre
f27	*Tag*	Pouvoir marquer un certain nombre de fichiers pour garder ou "créer" le lien archivistique	Fichiers	Fichiers tagués	
f28	*Sensivity review*	Identifier informations sensible dans les documents	Texte	Surlignage de données sensibles	
f29	Détection signature électronique	Identifier les signatures électroniques dans les métadonnées	Fichier	Métadonnées indiquant l'existence de la signature	ana_meta_signature
f24	Métriques archivistiques	Obtenir un "score" des DEVs de manière automatique	Métadonnées	Indicateur de confiance	
f30	Échantillonnage automatique	Échantillonner de manière automatique	Liste des fichiers à échantillonner	Liste des fichier échantillonnés	
f31	Renommage automatique	Si lot de données trop en vrac, permet d'imposer un système de nommage	Fichiers	Fichiers renommés	
f32	Réorganiser arborescence	Si lot de données trop en vrac, permet le calssement de celui-ci	Arborescence	Arborescence réorganisée	
f33	Rapport technique	Spécifier le versement ou l'élimination d'un élément du lot de données	Élément à modifier et indication de versement ou élimination (éventuellement commentaire sur la décision, date, etc.)	Élément modifié ainsi que tous les documents qui en dépendent hiérarchiquement	outil_modif_versement