

## Une petite histoire d'indexation alternative

**En préambule à cet article que je dédie au personnel de Fontainebleau et à Christiane et Gérard Naud qui m'ont fait aimer l'archivistique, j'émets le voeu qu'il ne soit pas sujet à polémique. C'est une réflexion toute personnelle laquelle est le fruit d'une longue expérience tant au sein des Archives territoriales que dans celles de l'État. Elle n'a pas vocation à remettre en cause la pratique de l'indexation mise en œuvre aujourd'hui qui reste pour moi un idéal plus que difficile à atteindre.**

En 2001, j'écrivais dans la Gazette des archives un article de 27 pages intitulé « *Indexation et analyse archivistiques, vers une recherche documentaire intégrale, de l'utilité des vocabulaires contrôlés a priori* » comprenant deux grandes parties : la première ayant pour objet « *L'indexation manuelle ou assistée par ordinateur : un outil secondaire, de l'analyse aux fiches cartonnées et électroniques* » ; la seconde « *L'indexation automatisée : un outil primaire, la simultanéité de l'analyse et de la recherche documentaire* »<sup>1</sup>. Quinze ans après cette contribution, je vais essayer d'en tirer les leçons et tâcher de voir si mon propos a encore du sens ou est devenu complètement obsolète. Inutile de vous dire qu'il n'a pas bouleversé la doctrine des archives et que le nombre de lecteurs n'a pas foisonné, et pourtant de prime abord il me semble contenir encore quelques vérités alors même que la pratique archivistique connaît de grands bouleversements.

Avant tout, il convient de définir l'indexation et de la replacer dans le contexte des grandes fonctions archivistiques. L'indexation est un élément de la description des documents/données et se traduit en une « *Opération destinée à représenter par les éléments d'un langage documentaire ou naturel des données résultant de l'analyse du contenu d'un document ou du document lui-même* »<sup>2</sup>.

C'est de l'indexation en langage naturel que traitera cet article<sup>3</sup>. L'une est manuelle et constituée de descripteurs, l'autre assistée par ordinateur et composée de mots-clés. Elles sont complémentaires mais ne connaissent pas les mêmes contraintes.

Il est utile ensuite de préciser que **l'indexation est à la source d'une révolution documentaire** entamée dans les années 1950 et 1960 avec les travaux de François-J. Himly<sup>4</sup> pour le Bas-Rhin puis Vital Chomel<sup>5</sup> pour l'Isère, et qu'elle a connu une forte accélération à partir des années 1970 avec l'effort de structuration de l'analyse archivistique marqué par les travaux de Christiane et Gérard Naud, avec l'utilisation aussi de bases de données pour la production des inventaires et comme instrument de recherche à part entière. Dans les années 1980, 1990 et 2000 s'épanouissent les grands chantiers normatifs nationaux<sup>6</sup> et internationaux<sup>7</sup> destinés à faciliter l'échange de données archivistiques dont la « toile » est devenue aujourd'hui le vecteur. Le développement des bases de données dans les services d'archives, leur mise à disposition dans les salles de lecture et l'accès qui en est offert à un public de plus en plus large par le biais de l'intranet et de l'internet, l'effort en matière de collecte et l'accroissement des collections qui en découle, l'archivage des données électroniques qui franchissent maintenant nos magasins et qui recèlent une documentation virtuelle très riche et toujours en mouvement, conduisent l'archiviste du 21<sup>e</sup> siècle à s'interroger sur l'adéquation de ses pratiques au contexte de production et de diffusion du matériau qu'il doit traiter.

**L'indexation est aussi le produit des vocabulaires normalisés.** Jusqu'à il y a peu, on considérait que les vocabulaires normalisés d'indexation demeuraient les seuls outils indispensables de la recherche documentaire. Pourtant, cette pratique n'est pas sans inconvénient pour le chercheur comme pour l'administrateur : abstraction du vocabulaire, diminution du champ sémantique, raccord artificiel de certains concepts aux descripteurs du thésaurus ou de la liste d'autorité.

---

1 *La Gazette des Archives*, n° 192, année 2001, pp 131 à 158.

2 *Dictionnaire de terminologie archivistique*, Direction des Archives de France, 2002 complété en 2007.

3 Pour l'indexation traditionnelle lire la contribution de Florence Clavaud dans « *L'Abrégé d'archivistique, principes et pratique du métier* ». Association des archivistes français, 3<sup>e</sup> édition revue et corrigée, Paris, 2012, pp. 180 à 187.

4 François-J. Himly a mis au point dès 1954 la forme la plus simple de langage documentaire permettant le regroupement des mots-matières de l'index par thèmes dans un tableau méthodique amorçant le rapprochement entre archivistique et techniques documentaires.

5 Vital Chomel publie un index des matières réalisé sur les versements administratifs.

6 Mise au point de thésaurus pour les matières, de listes d'autorité pour les typologies documentaires et les actions sur l'objet.

7 Publication des normes ISAD (G), ISAAR (CPF), ISDF, ISDIAH pour la description des documents, des producteurs, des fonctions et des institutions de conservation des archives par le Conseil international des archives.

Un temps même, on a cru pouvoir faire de **l'indexation un substitut de l'analyse**. Si elle « *reste un élément clé de la description [...] elle n'en est qu'un élément qui peut rendre compte de la valeur informative du document mais non de sa valeur probante* <sup>8</sup> ». Le langage naturel demeure le vecteur essentiel de la communication, celui qui permet la compréhension instantanée, le tri, la sélection et le choix. L'utilisation des langages documentaires arides et désincarnés place le lecteur en situation de traducteur qui s'efforce de restituer mentalement une structure afin de la rendre immédiatement intelligible.

On a pu croire aussi que **l'indexation pouvait être un complément de l'analyse** ainsi qu'aux vertus pythiques des descripteurs qui peuvent rendre fiables des analyses imprécises, élaborées et transcrites par les services versants dans les bordereaux de versement.

**L'indexation peut être aussi considérée comme une photographie de l'analyse.** La pratique actuelle de l'indexation et son étroite liaison avec les langages documentaires normalisés reste tributaire d'une tradition archivistique qui m'apparaît quelque peu anachronique. On raisonne en effet toujours en terme de mots-matières, et de noms propres (personnes physiques, personnes morales ou de collectivités, noms géographiques, titres d'oeuvres ou de journaux) c'est à dire en définitive en index *rerum, nominum* et *locorum* composant les inventaires analytiques traditionnels conformément aux circulaires diffusées entre 1841 et 1899 qui se fondaient sur une description à la pièce. Depuis, l'unité de description a changé au profit de l'article avec la création en 1909 du répertoire numérique et du répertoire méthodique en 1965, et d'autres niveaux comme le fonds, le sous-fonds, la série organique, le dossier qui peuvent s'articuler selon la norme internationale de description ISAD (G). Avec la puissance offerte par les technologies de l'information, le carcan des points d'accès imposés par la gestion des fichiers traditionnels a éclaté. Mesurer l'exactitude et la pertinence de la traduction documentaire du contenu de l'analyse implique donc aujourd'hui d'en privilégier les notions constitutives et désormais classiques<sup>9</sup>

<i>Corps de l'instrument de recherche</i>	<i>Zones de l'analyse</i>	<i>Éléments de l'analyse</i>	<i>Types d'indexation</i>
RÉFÉRENCE ARCHIVISTIQUE <sup>10</sup>		<b>Cote</b>	<i>Combinaison alphanumérique</i>
INTITULÉ (titre, analyse ou titre forgé)	<b>Auteur</b>	<b>Agent-auteur</b>	<i>Personne physique ou morale</i>
	<b>Sujet (titre)</b>	<b>Action</b> <b>Objet</b>	<i>Matière</i> <i>Matière, lieu-objet, personne physique ou morale</i>
		<b>Localisation dans l'espace</b> <b>Localisation dans le temps</b>	<i>Lieu localisant</i> <i>Matière, chronologie</i>
	<b>Pièces</b>	<b>Forme diplomatique</b> <b>Date intermédiaire</b> <b>Classement</b> <b>Lacunes</b>	<i>Matière</i> <i>Chronologie</i>
<b>Acteur</b>	<b>Agent-acteur</b>	<i>Personne physique ou morale</i>	
DATES EXTRÊMES		<b>Dates</b>	<i>Chronologie</i>

8 *La pratique archivistique française*, Direction des Archives de France, Paris, Archives nationales, 1993, page 146.

9 Naud (Christiane et Gérard), « L'analyse des archives administratives françaises contemporaines », dans la *Gazette des Archives*, n° 115, 4<sup>e</sup> trimestre 1981, pages 216 à 235. Doom (Vincent), « Description et analyse archivistique ou la nécessité d'une normalisation », dans la *Gazette des Archives*, n° 182-183, 3<sup>e</sup> et 4<sup>e</sup> trimestre 1998, pages 232 à 246.

10 Cet élément correspond au point d'accès 3.1.1., « Référence », de la norme internationale ISAD (G).

S'ajoutent à ce tableau les autres points d'accès des normes ISAD (G), ISAAR (CPF) et récemment ISDF et ISDIAH qui forment autant d'entrées et permettent en outre d'aller bien au-delà de la simple photographie laquelle n'est que la reproduction fidèle et besogneuse de l'analyse. L'indexation même assistée par l'ordinateur intégrant la gestion de thésaurus ou de notices d'autorité, reste une opération manuelle extrêmement fastidieuse et chronophage. Elle sécurise l'indexeur dont l'objectif a été, jusqu'il y a peu et le plus souvent, de produire des répertoires papier dotés de tables et qui refusait de considérer la base de données comme un instrument de recherche à part entière avec des spécificités échappant aux pratiques traditionnelles de l'archivistique.

**L'indexation automatisée est aujourd'hui devenue un instantané de l'analyse** grâce au développement de logiciels documentaires et de gestion électronique de documents et à la mise en œuvre d'une politique de records management. Pour ce faire, l'analyse se doit d'être dès l'origine pertinente c'est à dire définitive, complète, compréhensible par tous et dépourvue d'ambiguïté. Elle ne peut donc s'élaborer que dans le contexte de la collecte en associant les acteurs de la production détenteurs d'une véritable mémoire vive. Les tableaux d'archivage en forment les indispensables outils. Ils rendent aujourd'hui possible - à l'instar de la Bibliothèque nationale de France avec la base Rameau - la récupération à usage interne de notices d'autorité, sinon pointues du moins dotées des dénominateurs communs essentiels, le suivi de la description et de l'analyse d'une unité à l'autre, le traitement des suites documentaires fréquemment versées et leur pré-indexation. La mémoire vive et la conservation trouvent enfin l'équilibre indispensable qui permet de répondre à la pression du traitement en temps réel des fonds légitimement et universellement revendiqué. Cette démarche s'applique aisément aux archives anciennes et modernes.

Ainsi, l'indexation devient tout naturellement une opération parfaitement automatisée, simultanée de la saisie des données et notamment, mais pas seulement, du texte libre contenu dans l'intitulé au sens de la norme ISAD (G), qui peut comprendre un titre et/ou une analyse. L'indexation dite de « plein texte » ou « full text », ou encore de « texte intégral », est une indexation particulière dotée d'outils qui lui sont propres avec ses dictionnaires de mots vides, opérateurs de recherche et relations de synonymie et de hiérarchie dont l'organisation s'élabore *a posteriori*. Ainsi au fur et à mesure de la saisie des données, le vocabulaire se constitue, les mots vides sont retranchés, les relations sémantiques se construisent.

**Dans ce contexte, il convient de présenter le couple indexation intégrale et recherche documentaire.** La structuration de l'analyse qui sert de préalable permet de n'omettre aucun des éléments descriptifs essentiels à la transcription fidèle du contenu de l'unité et d'alléger par conséquent la grille de description de champs devenus redondants et qui alourdissent la saisie. La recherche documentaire en texte intégral dans une base de données archivistique repose sur un principe assez élémentaire : à une demande formulée à l'aide d'un terme issu du langage naturel, un mot-clé, ou d'une combinaison de termes, correspond une réponse constituée d'une ou plusieurs analyses ou descriptions structurées. Ce type de recherche utilise un langage de requêtes largement représenté sur les sites internet équipés de moteur de recherche et sur les logiciels de bases de données documentaires. Les requêtes sont formulées à l'aide d'opérateurs, les uns conventionnels, les autres adaptés au poids et à la diversité de l'information qu'ils recèlent :

<b>Opérateurs génériques</b>	EGAL	#	INFÉRIEUR OU ÉGAL	SUPÉRIEUR OU ÉGAL
<b>Opérateurs booléens</b>	ET	OU	SAUF	
<b>Troncatures</b>	@	*		
<b>Opérateurs de proximité</b>	PRÈS DE	« PHRASE »	« PARAGRAPHE »	« TITRE »

Les opérateurs de proximité sont à plus d'un titre les plus intéressants. Ils ont pour vocation de resserrer l'état de la recherche en la focalisant sur les différents éléments structurant les données. On pourra ainsi cibler la concomitance de deux termes ou plus au sein d'une même unité descriptive, et dans les unités les plus documentées au sein d'un paragraphe, d'une phrase, d'un titre ou de tout autre articulation prédéfinie. L'opérateur « PRÈS DE » permet, quant à lui, de recomposer des chaînes de caractères sémantiques. Il s'avère donc très utile pour la recherche dans les niveaux descriptifs élevés comme les sommaires des versements administratifs ou les descriptifs des fonds.

<i>Recherche</i>	<i>Opérateur</i>	<i>Interrogation</i>	<i>Résultat</i>
Journal La Croix du Nord	PRÈS DE	Journal PRÈS DE Croix PRÈS DE Nord	- [...] journal La Croix du Nord [...]
	ET	Journal ET Croix	- Croix en email ET journal intime de Sophie.
	OU	Journal OU Croix OU Nord	- [...] journal Nord-Eclair [...] - L'habitat minier dans le département du Nord [...]. - Inventaire des croix murales dans les églises du Douaisis [...].

Dans ce contexte, il apparaît donc particulièrement vain d'alimenter des champs dont le contenu figure explicitement dans l'analyse de l'unité décrite. Les noms propres en sont la parfaite illustration. Les archivistes focalisent souvent leur attention sur ces précieuses tables onomastiques. Et pourtant, quel intérêt y a-t-il à « copier » un terme, véritable mot-clé, pour le « coller » dans une rubrique constituant tel ou tel index. On objectera que l'orthographe d'un lieu varie souvent dans le temps, que les toponymes se succèdent les uns aux autres et s'enchaînent, que le bruit à la recherche sera important... A cela, il sera répondu que les relations de synonymie sont gérées dans un thésaurus, que les termes peuvent être hiérarchisés, que le bruit est atténué par les opérateurs de proximité qui complètent l'arsenal des opérateurs booléens classiques.

Les matières soulèvent à juste titre davantage de questions et d'interrogations. On limite dans un premier temps et sans aucune difficulté particulière l'imprécision de la recherche **en gommant les différences de genre et de nombre entre les termes** et en élaborant des relations de synonymie.

Dans un second temps, on peut constituer des hiérarchies entre les termes afin d'établir un véritable thésaurus, non plus de descripteurs mais de mots-clé, sans gêne aucune ni retard dans la mise à disposition des données. Ce thésaurus se caractérise alors comme un vocabulaire pragmatique élaboré *a posteriori* au fur et à mesure des entrées dans la base de données.

L'expérience montre toutefois que le nombre de termes issus du texte intégral est tel que cette opération devient une véritable gageure. Alors pourquoi vouloir aussi plaquer à un environnement éminemment évolutif un concept certes probant mais « archivophage », alors que d'autres éléments y pourvoient. La solution réside en grande partie dans la combinaison de critères présents dans les unités de description. On associera aux termes du texte intégral de l'analyse, d'autres points d'accès pour affiner la recherche comme : le service producteur dont la fiche descriptive apporte d'utiles renseignements ; mais aussi et surtout la fonction émanant de la norme ISDF.

On procède alors à une interrogation finalement assez proche des outils méthodiques inventés par François-J. Himly et Vital Chomel. Il paraît donc pertinent de réfléchir à une ou des listes d'autorité pour l'ensemble des services d'archives français qui puissent traduire de façon globale les grands domaines d'activité dont leurs fonds sont le témoin. Les cadres de classement des archives anciennes et modernes sont une piste intéressante pour les collectivités locales ainsi que l'instruction DAF/DPACI/RES/2009/018 sur le tri et la conservation des archives produites par les services communs à l'ensemble des collectivités territoriales (communes, départements et régions) et structures intercommunales. Les Archives nationales s'appuient depuis peu sur les fiches du plan d'orientation général (POG) qui couvre l'ensemble des domaines qu'elles couvrent. On pourrait aussi à l'instar des bibliothèques procéder à une indexation légère sur la notion de sujet qui, combinée au texte intégral de l'analyse, circonscrirait davantage la recherche.<sup>11</sup>

11 Grille de recherche d'un ouvrage dans la base de données OPSYS de la bibliothèque municipale de Douai.

<i>Critères</i>	<i>Données</i>
<i>Auteur</i>	Luxereau Anne
<i>Titre</i>	La malice de Monsieur Araignée : conte haoussa/Anne Luxereau ; éditeur (scientifique), Joëlle Busutil. - Paris : Gallimard, 1998. - 21 p. : ill. En couleur, (contes sans frontières ; 4)
<i>Sujet</i>	<i>Afrique : conte</i>
<i>Collection</i>	Contes sans frontières
<i>Editeur</i>	Joëlle Busutil
<i>Indice Dewey</i>	
<i>genre</i>	Conte
<i>résumé</i>	Une araignée africaine rend la vie impossible d'un village africain particulièrement difficile...

Quelles grilles de recherche doit-on en définitive retenir pour les archives ? La norme internationale ISAD (G) et ses 26 points d'accès associée à la norme ISAAR (CPF), à la norme ISDF et à la norme ISDIAH ainsi qu'au modèle d'analyse élaboré par Christiane et Gérard Naud correspondent pleinement à cet esprit. Il convient de définir non pas une grille, mais des grilles d'interrogation en fonction des utilisateurs : **un mode simple et un mode expert** où tous les points d'accès sont ouverts à la consultation. L'unicité de la grille internationale quel que soit le niveau de description ou la nature des unités renforce cette possibilité.

<b>Critères</b>	<b>Données</b>
<i>Provenance</i>	<i>Tous par défaut</i>
<i>Intitulé (titre et/ou analyse)</i>	<i>Théâtre (tous les fichiers par défaut)</i>
<i>Dates extrêmes</i>	<i>1980-1988</i>
<i>Fonction</i>	<i>Culture</i>
<i>Niveau de description</i>	<i>Tous par défaut</i>

**Dans une base de données archivistique les points d'accès permettant de localiser les unités documentaires dans le temps représentent des critères de choix ultimes et fondamentaux. Combinés à une recherche en texte intégral sur l'analyse qui elle-même contient certains éléments chronologiques complémentaires, ils limitent de manière incomparable les occurrences.**

Le public qui fréquente les services d'archives est de plus en plus au fait de l'évolution technologique et familiarisé avec l'informatique et les outils documentaires. On assiste en outre à un phénomène de plus en plus consumériste et de moins en moins érudit dans les méthodes de recherche. À nous de valider une information construite en amont, à la machine de permettre l'accès instantané en automatisant l'indexation dans l'attente d'une intelligence artificielle.

**L'indexation intégrale et l'édition** restent encore très liées selon que les services versants et les lecteurs disposent d'un outil informatique. Les bases documentaires intègrent toujours une fonction d'édition de tables annexées aux instruments de recherche. Le produit du traitement informatisé paraîtra, à n'en pas douter, particulièrement choquant pour les puristes qui vivent en chacun de nous. Des tableaux proches des outils méthodiques inventés par François-J. Himly et Vital Chomel pourraient être générés automatiquement et fondés non plus sur la sémantique mais sur le contexte mêlant les mots-clés quelle que soit leur nature.

<b>Fonctions</b>	<i>Administration communale</i>	<i>Pouvoirs du maire agissant au nom de l'État</i>	<i>Personnel communal</i>	<i>Urbanisme – voirie - communications</i>
<b>producteurs</b>	Cabinet du maire	Service de l'état civil	Service des ressources humaines	Service de la voirie communale
<b>Mots-clés</b>	correspondance	Décès	Debrabant, Georges	Émile-Zola
	Délibération	Dorignies	Notation	Dénomination
	Maire	Mariage	rémunération	Rue
	Vernier, Jacques	Naissance	Vaccination	ZAC

Enfin, **l'indexation intégrale et la description rétrospective** sont bien prises en compte que ce soit pour les archives électroniques, les inventaires anciens impossibles à saisir ou à reprendre compte tenu de leur volume, et les instruments de recherche – bordereaux de versements et répertoires – élaborés à l'aide de la bureautique. Il y a peu encore la numérisation avec reconnaissance optique de caractère (OCR) a représenté une alternative intéressante. Désormais la dématérialisation se fait par encodage en EAD/XML mais là encore, il est nécessaire de poser des balises en amont.

En définitive, mon propos conduit à affirmer avec force, s'il en était besoin, que l'indexation, comme l'archivistique en général, repose sur la description et l'analyse et qu'elle doit faire figure d'opération transparente et entièrement automatisée, autorisant ainsi le transfert d'un temps archivistiquement précieux là où il est le plus nécessaire, en amont du processus de constitution des fonds. On en revient inmanquablement à la théorie des flux et au *records management* qui contraignent à mettre en œuvre une politique active auprès des différents acteurs de l'archivage. Les chartes d'archivage, contrat moral et quasi juridique, établies par fonction entre les producteurs et les conservateurs, en forment un excellent préalable<sup>12</sup>.

Les outils documentaires élaborés par la profession ne sont pourtant pas à mettre au panier, loin s'en faut. Les thésaurus et listes d'autorité sont en effet indispensables à l'analyste pour qualifier la richesse des typologies documentaires, des formes externes, des techniques et des supports, pour structurer in fine la matière et les éléments onomastiques constituant le dictionnaire du texte intégral. Les Archives nationales ont d'ailleurs opté pour les deux types d'indexation : manuelle et automatisée en fonction des besoins. Un grand pas a été franchi tout naturellement avec l'indexation intégrale mise en œuvre dans la salle des inventaires virtuelle. Elle n'est pas exclusive comme le démontre le minutier central des notaires. Cette avancée vers l'automatisme est me semble-t-il inéluctable et n'est nullement une solution d'attente comme le pense nombre d'archivistes<sup>13</sup>. On assiste à l'émergence d'une intelligence artificielle. Le développement de l'information dématérialisée et polymorphe conduit donc à privilégier la structuration des données et leur encodage.

**Alors, consacrons dès à présent notre temps et notre énergie à la description des unités documentaires, issue du matériau primaire, plutôt qu'à l'indexation, son dérivé, et participons à la constitution d'un système automatique pertinent. C'est là une question de stratégie compte tenu du volume des archives à traiter.**

## Vincent Doom

12 Voir la charte d'archivage comprenant le tableau de tri et de conservation des archives de la préfecture du Nord, Lille, 2005.

13 D'aucuns me disent que les Archives nationales sont dotées de plusieurs bases de données indexées manuellement comme le minutier central des notaires. À cela, je répondrai quelle est la part des fonds indexés ? Enfin, les catalogues des plans de la série F des Archives nationales n'ont pour le moment aucune indexation.