PUBLIC INFO PRESSE SERVICE

LA DOCUMENTATION DE PRESSE AUTOMATISÉE À LA BIBLIOTHÈQUE PUBLIQUE D'INFORMATION

ANNE GOURHAND
CLAIRE STRA
Bibliothèque publique d'information

Public Info, service de documentation destiné aux utilisateurs de la Bibliothèque publique d'information, propose un fonds de dossiers de presse qui permet à chacun de retrouver des documents sur des sujets d'actualité culturelle et sociale qui n'ont pas encore fait l'objet de monographies.

Communication et conservation

Le projet d'installation d'un système de gestion électronique de ce fonds a pris en compte deux problèmes de fonctionnement : la communication manuelle de documents séparés, fragiles et de tailles variables et leur récupération, la conservation et le stockage dans un espace restreint d'un fonds en accroissement.
La vocation de la BPI étant d'expérimenter de nouveaux outils technologiques, l'installation d'un système GED (gestion électronique des documents) dans le contexte du Centre Georges-Pompidou présentait un intérêt d'usage non négligeable.
En effet, alors que cette technologie a déjà une existence dans d'autres cadres professionnels, elle n'avait pas encore été testée dans un environnement grand public, sur des postes autonomes.
Par ailleurs, l'objectif de l'installation d'un système de production et d'indexation de documents numérisés était d'adapter à une équipe de documentalistes travaillant de manière collégiale un outil de travail automatisé, dont la souplesse d'utilisation sauvegarde l'organisation du travail en concertation, et permette de contrôler, sans la bloquer, la cohérence intellectuelle du travail documentaire. Ce contexte est notablement différent de celui des applications bureautiques de la GED, au sein duquel les tâches des équipes de production sont très compartimentées, attribuées à différentes catégories d'agents, et gérées par un administrateur unique.
C'est ainsi qu'a été initié ce projet, avec une dotation budgétaire moyenne et une priorité assignée à la recherche de logiciels standard, paramétrables aux besoins de l'application, en limitant au maximum les travaux coûteux de développements spécifiques.

Le fonds documentaire et sa gestion

Il s'agit d'articles issus de la presse française d'information générale, sur l'actualité culturelle et sociale. Cette base est organisée en deux parties : des dossiers thématiques sur les grands sujets d'actualité ; des dossiers biographiques, consacrés à des personnalités du monde entier (artistes, écrivains, metteurs en scène, comédiens, intellectuels, etc.).
Ce fonds, constitué de quelque 300 000 pages A4, s'accroît d'environ 10 000 documents nouveaux par an. La partie thématique est constamment épurée, afin de correspondre à trois

années d'information. Créée en 1977, la partie biographique s'enrichit d'année en année. Une équipe de dix bibliothécaires/documentalistes gère cet ensemble, et assure indifféremment toute la chaîne de traitement du document, qu'il soit manuel, comme naguère, ou automatisé, comme aujourd'hui. Les travaux quotidiens

LE CHOIX DU RÉSEAU EST CELUI D'UNE ARCHITECTURE CLIENT/SERVEUR PERMETTANT UNE MISE À JOUR SIMULTANÉE DES POSTES DE GESTION ET DES POSTES DE CONSULTATION

doivent pouvoir être interrompus par les documentalistes autant de fois qu'il est nécessaire pour assurer le service prioritaire qu'est l'accueil sur place des utilisateurs. Rappelons que la BPI accueille en moyenne 10 à 12 000 personnes par jour, six jours – soit soixante-quatre heures – par semaine, dimanches et jours de fêtes compris.

Caractéristiques techniques des outils choisis

Le choix du réseau est celui d'une architecture client/serveur permettant une mise à jour simultanée des postes de gestion et des postes de consultation, ceci afin de préserver la rapidité d'accès de l'utilisateur final à l'information (fonds d'actualité). Ont été mis en place un système de sauvegarde des données très sécurisé par l'utilisation d'un serveur équipé d'un système RAID 5, un câblage de fibres optiques, des postes de travail internes équipés de scanners de bureau de format A3, des postes publics avec des écrans tactiles et un système de paiement des impressions de documents par carte monétique. Cet équipement a été mis en œuvre par la société SIATEL.

Deux modules de traitement des documents ont été sélectionnés : les logiciels FULDESK et FULTHES de la société FULCRUM, respectivement gestionnaires de bordereaux d'indexation, de listes de références pour le premier, et d'un module de thésaurus, pour le second. Ces outils ont été paramétrés sur la base d'un modèle conceptuel des données (MCD) documentaires de Public Info, réalisé par le service avec l'aide d'un expert extérieur.

L'organisation du travail

Un des aspects exemplaires de cette réalisation a sans doute été la mise en place par les documentalistes d'instruments de travail strictement adaptés au traitement de ce fonds, à partir des progiciels choisis, en particulier le thésaurus et les listes contrôlées complémentaires. La difficulté était d'adapter à une organisation du travail très collégiale, et qui a fait ses preuves en vingt ans, un outil automatisé cohérent, sans avoir recours à des développements de programmes spécifiques.

Il s'agissait aussi de respecter la sémantique du fonds documentaire en créant une structure de travail qui permette de contrôler l'indexation et d'en refléter strictement le contenu. Il semblait irréalisable d'adapter un outil déjà existant. La preuve a été faite qu'il était possible d'instaurer un système de contrôle souple et fiable, sans rigidifier ni ralentir le travail de traitement des documents. Les caractéristiques techniques du logiciel choisi ont facilité cette mise en œuvre.

Le noyau de FULDESK est fondé sur le principe du *full text,* mode d'indexation et de recherche qui permet de retrouver un document par un ou plusieurs mots du texte ou d'un champ du formulaire d'indexation associé à celui-ci. Il peut donc gérer de nombreuses listes de référence spécifiques, contrôlées ou non. Celles-ci peuvent être aisément enrichies de données nouvelles, avec mise à jour simultanée dans l'architecture client/serveur.

Mise au point de la méthodologie

La fin de l'année 1994 a été consacrée à une phase de réflexion théorique et d'analyse du fonctionnement du service. Des groupes de travail se sont répartis l'élaboration d'outils normatifs : manuels de catalogage et d'indexation des articles, normes de choix des documents, réflexion sur la méthode de construction du thésaurus.

L'accès à la documentation se faisait jusqu'alors par l'intermédiaire de fichiers manuels rotatifs : un fichier général de données classées dans l'ordre alphabétique de mots matières, inspirés au départ de la liste-autorité de la BPI, et des fichiers signalétiques satellites reliés par des renvois au fichier principal.

Nous avons pu constater, à partir de l'ouverture du service au public en 1988, un assouplissement de l'indexation, la multiplication de renvois et de descripteurs plus proches du langage naturel. Cette pratique était liée à l'introduction dans l'équipe d'accueil du public de personnes extérieures au service, qui devaient pouvoir utiliser cet instrument sans intermédiaire.

Cette modification d'approche, la nécessité d'actualité de l'indexation et de son introduction rapide dans les fichiers ont déterminé notre choix entre deux modèles conceptuels de données. Le MCD, fondé sur la construction de vedettes matières d'indexation, a été écarté en faveur d'un modèle conceptuel de données établi sur la notion de descripteurs d'indexation.

Pour des raisons de capacité de gestion, l'élaboration d'un seul thésaurus ne pouvait suffire à prendre en

Abréviations

Champ sémantique	CS
Descripteur	DES
Domaine	DO
Synonyme	SY
Terme associé	TA
Terme éphémère	TE
Terme générique	TG

compte tous les descripteurs nécessaires à la future base. Le module FULTHES retenu pouvant gérer jusqu'à 10 000 descripteurs (6 000 de façon optimale), nous avons décidé, par économie, et en exploitant les capacités du moteur FULCRUM, de lui adjoindre :
– une liste contrôlée de noms propres et d'organismes ;
– une liste contrôlée de titres d'œuvres ;
– une liste de termes « éphémères » (TE), constituée au fur et à mesure par les indexeurs, regroupant les expressions de l'air du temps souvent utilisées par le public pour interroger, mais que l'on ne souhaite pas intégrer au thésaurus. Ces termes sont ajoutés à l'indexation classique d'un article, par exemple, *Eurosceptique* (TE), ajouté aux descripteurs *Union européenne* et *Opposition*.

Méthode de construction du thésaurus

L'orientation essentielle de la réflexion était de partir avant tout de notre documentation, afin que le vocabulaire retenu soit le plus proche possible des termes couramment utilisés dans la presse généraliste.
Plusieurs thésaurus ont été examinés *(Le Monde*, BIPA-Banque d'information politique et d'actualité de la Documentation française, Mémobase du Centre régional de documentation pédagogique de Poitiers). Celui de la BIPA s'étant révélé le plus proche de notre fonds, il a été utilisé comme canevas pour effectuer un premier tri dans la masse des dossiers. La documentation (environ 4 000 dossiers) a été répartie en dix-sept grands sujets reprenant les domaines (DO) et les champs sémantiques (CS) de la BIPA qui convenaient, éliminant les sujets non traités à Public Info (par exemple, les entreprises, les finances publiques, l'industrie et l'économie générale) ou regroupant des domaines s'ils étaient jugés trop larges pour notre usage (par exemple, *Institutions* et *Vie politique)*. Par la suite, nous avons appliqué en l'adaptant la méthode classique d'élaboration d'un thésaurus : collecte, épuration, ventilation, structuration.

L'ORIENTATION ESSENTIELLE DE LA RÉFLEXION ÉTAIT DE PARTIR AVANT TOUT DE NOTRE DOCUMENTATION AFIN QUE LE VOCABULAIRE RETENU SOIT LE PLUS PROCHE POSSIBLE DES TERMES COURAMMENT UTILISÉS DANS LA PRESSE GÉNÉRALISTE

L'ensemble du service – par groupes de deux ou trois personnes – devait collecter les termes de la totalité des dossiers de presse (documentation biographique exceptée) en se répartissant les différents domaines. Plusieurs objectifs complémentaires ont été fixés : noter les termes ou expressions qui reviennent dans les sommaires des dossiers et dans les coupures elles-mêmes, penser aux termes qu'utiliserait le public pour sa recherche plutôt qu'aux termes satisfaisants pour la structuration, retenir les notions réellement significatives – c'est-à-dire pour lesquelles il n'existe pas moins de cinq articles. Pour aider à organiser cette collecte, nous avons mis au point des grilles de saisie, déjà structurées, simplifiées par la suite – il était plus pratique de noter les termes comme ils se présentent que de vouloir commencer la structuration.
L'équipe a pris la décision de confier à un de ses membres la tâche de coordonner le travail d'élaboration. En effet, la nécessité d'un seul interlocuteur pour tous les groupes de collecte s'est rapidement fait sentir. Il fallait assurer la cohérence du travail, avoir une vision d'ensemble permettant d'harmoniser le vocabulaire choisi, coordonner le niveau d'arborescence (deux niveaux de termes spécifiques au maximum), et éviter les redondances. L'objectif prioritaire était de construire un outil destiné au grand public et non d'être aussi précis qu'un thésaurus scientifique.

Élaboration du thésaurus

L'élaboration du thésaurus s'est déroulée en dix-huit mois et plusieurs étapes. La collecte a été effectuée par toute l'équipe en deux ou trois séances par semaine pour chacun des participants, en fonction des disponibilités laissées par l'accueil du public et les tâches quotidiennes.
Chaque groupe devait collecter les termes de tous les dossiers d'un même domaine, un champ sémantique après l'autre, par exemple les trente-trois dossiers du CS *Environnement* du DO *Cadre de vie*. Cette méthode permet de balayer de grands pans de documentation afin de repérer le vocabulaire commun à plusieurs branches et sous-branches[1].

1. *Acteur, metteur en scène, habilleur,* sont rassemblés sous le CS *Métiers du spectacle* plutôt qu'annexés à la seule branche *Théâtre.*

Les séances de ventilation-structuration des termes collectés se sont faites avec le coordonnateur qui, pendant ces dix-huit mois, s'est consacré en priorité à l'élaboration du thésaurus. Il s'agissait tout d'abord de répartir les termes collectés par groupes de même famille, de repérer les termes génériques (TG) qui se dégagent, puis de lister les termes spécifiques (TS) qui en dépendent. Lors de cette opération s'effectuait également le choix des termes retenus et des termes rejetés en synonymes (SY).

Ce travail de répartition a mis en lumière des problèmes de plusieurs types :

– *le niveau de structuration des branches*. Des CS se sont révélés trop importants pour rester sous leur DO et ont dû être remontés d'un cran dans l'arborescence. C'est le cas du DO *santé* à l'origine sous *questions sociales*, ou du DO *religion* sous *culture* (dans le thésaurus de la BIPA). Le nombre des domaines est ainsi passé de dix-sept à vingt-cinq ;

– *le type de structuration*. L'analyse de la structure d'un sujet a été préférée au regroupement des familles sémantiques, tout en assurant la cohérence par les passerelles que sont les termes associés (TA)[2]. Ce type d'organisation ouverte offre une plus grande souplesse pour le rattachement de nouveaux termes et limite les risques de redondance. De même, il permet d'associer un organisme donné à différents sujets au lieu d'en faire le terme spécifique d'un seul ;

– *l'harmonisation des niveaux de langage* entre les différents domaines. La préférence a été donnée au terme courant utilisé par la presse ;

– *la ventilation des termes* appartenant à plusieurs domaines et/ou jugés inclassables dans un premier temps. L'avancement dans l'élaboration du thésaurus a montré qu'ils s'imbriquaient finalement assez facilement dans l'arborescence, justifiant en quelque sorte *a posteriori* la cohérence des choix de structure décrits plus haut.

Entrée des données dans le module FULTHES

En ce qui concerne la saisie des termes, ce gestionnaire de thésaurus est d'une grande simplicité d'utilisation. Il se présente sous la forme d'un cadre à remplir divisé en deux parties : un *En-tête* contenant le descripteur et son environnement sémantique (terme générique et rappel des DO et CS concernés) et un cadre *Relations* avec des fenêtres de saisie pour les TS, les SY, les TA ainsi qu'une zone de notes d'application. Les corrections sont faciles à effectuer et le contrôle exécuté sur les doublons par le programme s'est avéré très pratique en cas de terme déjà intégré dans un autre domaine.

La saisie se fait sur le domaine déjà structuré, cette méthode s'est révélée plus simple que de saisir à la chaîne les candidats descripteurs dans le thésaurus et de les raccrocher directement dans les différentes branches. Elle offre une meilleure vision d'ensemble. Des modifications de structure ont parfois été apportées au moment de la saisie, car le module permet de mieux visualiser les différentes branches et sous-branches et de rééquilibrer des termes génériques ou spécifiques en les faisant remonter dans l'arborescence.

Validation, corrections

Le début de la numérisation des articles, en septembre 1995, a permis de tester en « vraie grandeur » la validité de cette méthode et d'y apporter quelques modifications. Le nombre de termes entrés en octobre 1995 – 2 600, soit moins de la moitié de notre quota maximum par rapport au nombre de domaines déjà traités (plus de la moitié) – a autorisé un peu plus de souplesse dans la construction, c'est-à-dire moins de rejets de termes en synonymes. Dans le même sens, la pratique de l'indexation nous a amenés à transformer des synonymes en termes spécifiques, princi-

2. Par exemple, dans le CS *transports*, ont été choisis les TG *Infrastructure des transports (aéroport, gare, route...), Moyen de transport (automobile, avion, bateau...)*, plutôt que les TG *Transport ferroviaire (gare, train, rail...)* et *Transport aérien (avion, aéroport...)*.

palement dans deux cas : selon les besoins de l'actualité[3], et sur demande des indexeurs qui ne sont pas satisfaits de l'association du terme retenu avec d'autres descripteurs[4].

Procédures de révision

La phase de construction initiale du thésaurus s'est achevée en juin 1996. A cette date, les neuf mois de numérisation effective des articles de la base ont permis de constater le bon fonctionnement du système, c'est-à-dire de l'indexation des articles par l'intermédiaire du thésaurus et des listes.

Une nouvelle phase a pu alors commencer : celle de la révision complète du thésaurus, travail de longue haleine de relecture et d'harmonisation effectué domaine par domaine ; celle – très importante au quotidien – des corrections et des ajouts demandés par les indexeurs au moyen d'un cahier de bord.

Plusieurs solutions sont apportées en fonction du contexte : la création d'un nouveau descripteur inséré dans son domaine, le décrochage de synonyme transformé en terme spécifique, le refus de création correspondant généralement à une indexation trop précise, la proposition de créer le terme dans la liste des termes éphémères ou celle des noms de personnes.

Dernier cas – le plus délicat –, la correction de descripteurs (nouvelle appellation d'un organisme, par exemple) ou la transformation de termes éphémères en descripteurs qui nécessitent une modification de la base articles, soit globale si elle concerne beaucoup d'articles, soit notice par notice.

En janvier 1997, le thésaurus comptait 5 186 descripteurs. Ce travail d'enrichissement constant, en liaison étroite avec l'indexation des articles, est essentiel pour faciliter l'accès de l'utilisateur final au document en lui offrant un outil en perpétuelle évolution.

L'interface publique

La mise en œuvre de l'organisation du travail décrite plus haut avait pour but ultime la création d'une interface publique la plus simple et la plus intuitive possible. Le lecteur devait pouvoir naviguer dans la base documentaire avec les termes utilisés tous les jours dans les médias, et récolter le maximum de documents pertinents, sans avoir à créer une équation de recherche complexe. Rappelons que les postes de travail sont mis à disposition des usagers, sans que ceux-ci aient une formation préalable, la logique d'utilisation doit donc être évidente.

L'accès aux documents numérisés est unique, par mots-clés, ceux-ci pouvant être un nom commun, une expression usuelle, un nom propre, un titre ou les mots d'un titre, un nom d'organisme, une date, un sigle, un nom géographique, etc., sans contrainte particulière d'écriture et sans opérateurs de recherche explicites.

Le système gère de manière transparente les opérateurs booléens, les singuliers/pluriels réguliers ou irréguliers, ainsi que les synonymies et les termes associés mis en place dans le thésaurus. Dans les cas de recherche infructueuse, il assure aussi une extension automatique de celle-ci sur les termes associés, et sur les termes génériques pour les noms géographiques, le but de cette organisation étant, bien entendu, de limiter au maximum les silences.

L'utilisateur parvenu à la consultation d'un document bénéficie aussi d'un affichage actif de tous les mots-clés choisis par les documentalistes pour indexer celui-ci, et peut initier une nouvelle recherche immédiatement à partir de l'un de ces mots. Il navigue ainsi très simplement dans le fonds documentaire.

Un second accès a été ménagé, qui permet au public de consulter des ensembles documentaires déjà organisés en amont, sortes de « dossiers électroniques », sur des thèmes d'actualité très brûlants – par exemple, *Crise de la vache folle, Question de l'immigration clandestine,* etc. La revue de presse est ici toute faite, au lieu d'être constituée par l'utilisateur lui-même.

Les premières observations de l'usage des postes publics montrent que les utilisateurs sont très rapidement à l'aise dans la consultation, sans préparation particulière. L'interface de production a, elle aussi, été rapidement assimilée et maîtrisée par les documentalistes : opérations de numérisation et d'indexation avec l'aide du thésaurus et des listes de référence, dans la mesure où l'équipe avait été complètement associée à l'élaboration et aux spécifications des outils automatisés.

Il n'y a donc pas d'incompatibilité à mettre en place de tels instruments de travail au sein d'une organisation très collégiale, si ceux-ci présentent une réelle flexibilité fonctionnelle. Ceci est valable pour l'élaboration de la structure finale de recherche : l'interface publique de la base documentaire.

Il faut prendre en compte cependant deux difficultés permanentes : le suivi technique essentiel pour l'avancement harmonieux de tels projets reste trop ponctuel dans la plupart des cas, pour permettre d'éviter les difficultés d'interprétation et de traduction des besoins professionnels par les sociétés de service informatiques.

Il ne faut pas non plus minimiser le temps nécessaire à la réelle mise en œuvre de tels systèmes. Une analyse rigoureuse des données à traiter, des usages de production et d'utilisation de celles-ci, est nécessaire à l'élaboration d'un produit exactement adapté aux besoins mis en évidence.

Février 1997

3. Par exemple, *maladie à prion* = DES (descripteur), *maladie de Creutzfeldt-Jacob, maladie de la vache folle, tremblante du mouton* entrés d'abord en synonymes, puis sortis en termes spécifiques.

4. Par exemple, le DES *Inondation (Boue* = SY) devient incongru pour l'indexation d'un article sur le recyclage des boues d'épuration.