Anne-Marie Motais de Narbonne
Direction de l'information scientifique et technique et des bibliothèques

PANCATALOGUE

UN CATALOGUE COLLECTIF DE LIVRES POUR L'ENSEIGNEMENT SUPÉRIEUR

PANCATALOGUE est le catalogue collectif des ouvrages acquis par les bibliothèques des universités et des grands établissements. Sa finalité première est la localisation des ouvrages en vue du prêt entre bibliothèques mais, en outre, par le nombre des documents recensés – plusieurs millions –, il constituera aussi un outil de recherche bibliographique. Sa conception a posé de nouvelles questions sur les données bibliographiques, les données de localisation et sur leur organisation dans un ensemble cohérent. Les volumes de données à traiter, le mode d'alimentation par chargement, l'hétérogénéité des notices d'origine, élaborées selon deux règles de catalogage (Afnor et AACR2) et présentées sous deux formats différents (USMarc et Unimarc) se sont conjugués pour imposer des choix originaux. Le projet initialisé en 1987 a déjà fonctionné dans deux versions distinctes. Il est parvenu aujourd'hui à maturité avec une base de plus d'un million de notices, alimentée par trois types de données avec la participation de quelque 80 établisements. C'est l'histoire de cette réalisation qui est racontée ici.

Contexte général et genèse du projet

Au-delà de tous les changements de structures administratives et de techniques disponibles, parce que c'est un besoin fort de la recherche, le ministère en charge des bibliothèques universitaires a toujours porté un intérêt majeur à leur coopération. Les réalisations en ce sens ont en outre été facilitées par le mode de gestion longtemps très centralisé de ces bibliothèques qui, s'il présenta par ailleurs de sérieux inconvénients, n'en a pas moins contribué à les doter de méthodes communes et d'unetradition de travail en réseau.
C'est dans cette continuité qu'à partir des années 1980, l'informatique a été utilisée pour créer le catalogue collectif des périodiques (CCN), le prêt entre bibliothèques (PIB puis PEB), le recensement national des thèses (Téléthèses), et enfin le catalogue collectif des ouvrages dont il est question ici.
Cette séquence des réalisations informatiques reflète à la fois l'ordre des priorités à satisfaire et celui imposé par les contraintes techniques du moment. En effet, le prêt entre bibliothèques est l'aboutissement concret de toute la politique de coopération, la première raison d'être des catalogues collectifs et plus de 90 % de cette activité porte sur des périodiques. Par ailleurs, au plan organisationnel, un catalogue collectif des périodiques pouvait être réalisé de façon autonome, sans toucher au fonctionnement courant des bibliothèques participantes.
C'est dans ce contexte et pour contribuer au même objectif de rendre accessible à tous les chercheurs l'ensemble des ressources documentaires des universités qu'en 1987 la Direction des bibliothèques, des

musées et de l'information scientifique et technique (DBMIST) lançait le projet de catalogue collectif des livres et initialisait, pour les bibliothèques universitaires, une politique nationale de catalogage normalisé en réseau qui allait en garantir l'alimentation.

Les principes d'alimentation

L'ampleur du projet par le volume des données et le nombre des institutions participantes est un élément clef pour les solutions qu'il s'agissait de mettre en place. En effet, on compte plus de 80 bibliothèques universitaires comprenant environ 300 sections géographiquement séparées. On estime à 400 000 le nombre de leurs acquisitions annuelles et leurs fonds à 21 millions de livres. Ces quantités sont déterminantes pour tous les choix. Elles imposent, par exemple, que tous les traitements soient automatiques et, au plan organisationnel, elles interdisent que des travaux de catalogage soient faits en double. Le catalogage initial doit donc impérativement servir de base à l'alimentation du catalogue collectif, sans intervention manuelle supplémentaire.

La solution informatique classique est de constituer un catalogue collectif directement dans la base de catalogage partagée en assimilant en quelque sorte l'établissement qui crée ou utilise la notice avec l'établissement où le document est localisé. Le catalogue collectif est alors, techniquement, un produit dérivé du catalogue partagé.

Une telle mise en œuvre supposait l'existence d'un dispositif de catalogage partagé capable de supporter la totalité des volumes à traiter. Aucune infrastructure de cette dimension n'existait alors en France et il n'était pas possible d'en créer sans délais importants. En outre, une fois un tel dispositif mis en place, il aurait encore fallu beaucoup de temps pour obtenir les effets attendus, tant que n'aurait pas été constituée une masse suffisante de données.

Un catalogue collectif indépendant des structures de catalogage, alimenté par des données extérieures, s'imposa donc, à l'époque, comme la seule voie pour obtenir dans des délais raisonnables une base utile au prêt entre bibliothèques. La question se posait alors de l'origine des données.

Depuis 1985, les bibliothèques universitaires disposaient d'un système de catalogage sur micro-ordinateur appelé Mobicat, permettant de produire des données informatiques et des fiches. Les travaux préparatoires à un projet commun avec le ministère de la Recherche appelé CCO (catalogue collectif des ouvrages) avait révélé la forte hétérogénéité de ces données et condamné leur utilisation pour la constitution d'un catalogue collectif commun à l'ensemble des universités.

Un catalogue collectif indépendant des structures de catalogage, alimenté par des données extérieures, s'imposa

En revanche, la Bibliothèque nationale annonçait que les données de la Bibliographie de la France seraient prochainement disponibles, celles de la base Sibil-France l'étaient déjà et d'autres bibliothèques pouvaient commencer rapidement à travailler avec l'organisation nord-américaine de catalogage partagé OCLC. Cet environnement conduisit à deux décisions complémentaires et indissociables. Pour les bibliothèques, le principe fut arrêté du catalogage normalisé dans un réseau de catalogage partagé et, pour le catalogue collectif, celui d'une alimentation par le chargement des données correspondantes. Les bibliothèques participantes au catalogue collectif eurent donc l'obligation de réaliser leur catalogage dans une des trois sources reconnues par le dispositif national, Sibil-France (Montpellier), OCLC (Dublin, Ohio), et BN-Opale (Paris), le choix de la source appartenant à chaque université et des accords spécifiques étant conclus pour la livraison directe des données par les sources au catalogue collectif.

Le problème de la fusion des données dans une base unique était ainsi réduit au traitement de trois sortes de notices, certes élaborées selon des principes différents, mais dont la normalité garantie assurait la faisabilité de l'opération. C'est donc sur ces principes que le projet de catalogue collectif des ouvrages appelé Pancatalogue était lancé.

La phase d'initialisation

Ces principes établis, commença la phase de lancement et d'initialisation devant conduire, dans un premier temps, au choix des logiciels puis à leur adaptation aux besoins spécifiques de l'application.

Le choix des logiciels

Les promoteurs du projet étaient bien conscients du caractère original de l'entreprise et des longs développements qui seraient nécessaires à son aboutissement. Cependant, pour éclairer les choix informatiques fondateurs, il s'est pourtant agi, en premier lieu, de définir le projet dans sa totalité en précisant les objectifs, les fonctionnalités, les traitements de données, les volumes d'activité, les contraintes, etc. Tous ces points furent étudiés par une équipe de projet avec l'aide d'un groupe de travail composé de directeurs de bibliothèques et d'informaticiens. L'ensemble des spécifications arrêtées fut formalisé dans un document de plus de 50 pages intitulé *Cahier des charges du Pancatalogue,* système de gestion et de consultation d'un catalogue collectif national d'ouvrages qui constitua le document technique des appels d'offres.

En décembre 1987, la première étape de la phase d'initialisation du projet s'achevait avec le choix, parmi cinq offres concurrentes, du logiciel Dobis Libis (IBM), et d'un logiciel pour le

chargement des données OCLC qui avaient été choisies pour initialiser l'application.

Les principes de chargement et les données de localisation

Les données à charger se présentent comme une succession de notices bibliographiques comprenant aussi des indications de localisation. Pour obtenir un regroupement de l'ensemble des localisations d'un document sous la notice qui le décrit, les chargements doivent être différenciés selon les cas. Si le document n'existe pas encore, il s'agit de créer une nouvelle notice avec une première localisation ou, si la notice existe déjà, il ne faut créer que la nouvelle localisation. Enfin, si notice et localisation figurent déjà, il faut rejeter les données. C'est donc sur la définition des identifiants de notice et de localisation, sur leur reconnaissance dans les données d'entrée et sur leur comparaison avec leurs homologues déjà présents dans la base que repose le pilotage des chargements.

Les identifiants de notices sont les numéros attribués par les sources d'origine et conservés dans les notices Pancatalogue. En revanche, le choix du système d'identification des localisations est plus délicat.

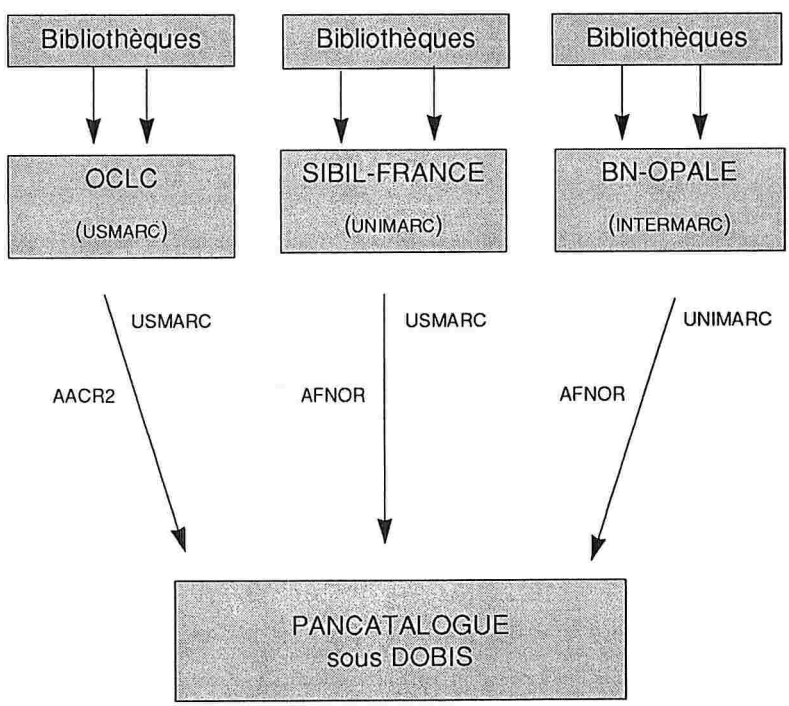
La première cause de complexité tient, outre tous les problèmes de codification, à la structure des données de localisations. L'institution propriétaire du document et son lieu de conservation, différents l'un de l'autre, constituent deux intitulés, liés le plus souvent de façon hiérarchique. Suivant l'organisation du service, la demande de prêt entre bibliothèques doit être faite au premier intitulé ou au second, voire à une troisième adresse. Les données sur les exemplaires comme les cotes ou les numéros inventaires n'ont de sens que liés implicitement ou explicitement au lieu de conservation et constituent ainsi un troisième niveau hiérarchique. Les notions de fonds peuvent également intervenir selon une définition géographique et locale ou une définition thématique et constituer un niveau supplémentaire. On voit donc que, pour les données de localisation, cinq éléments avec quatre niveaux hiérarchiques au moins seraient nécessaires, alors que les possibilités offertes par les systèmes informatiques de gestion sont en fait bien plus réduites.

Cette situation générale à laquelle s'ajoute l'absence complète de normalisation pour ces données a conduit les organisations de catalogage à faire leurs propres choix de structure. Pancatalogue en est dépendant à travers les données qu'il reçoit et il doit aussi prendre en compte les contraintes du logiciel Dobis et les besoins du prêt entre bibliothèques.

Si Dobis permet trois niveaux de descriptions, institutions, succursales et exemplaires par exemple, la hiérarchie réelle est organisée sur deux niveaux. Ainsi, deux exemplaires d'un même document dans deux institutions Dobis sont gérés séparément, (les institutions étant autonomes en matière d'acquisitions et de politique de prêt), tandis que deux exemplaires de deux succursales font partie de la même unité de gestion. Pour le prêt entre bibliothèques, il est souhaitable que les localisations soient différenciées au niveau des sections puisqu'à celles-ci correspondent des conditions différentes d'accessibilité. Le dispositif de chargement doit donc accepter deux notices identiques avec des mentions de sections différentes, mais rejeter deux notices identiques provenant d'une même section.

L'arbitrage définitif entre toutes ces contraintes est concrétisé dans l'installation des *tables* du « réseau Dobis-Pancatalogue », sorte de déclaration préalable des localisations reconnues par le système qui constitue le fichier

Pancatalogue
Mise à jour de la base



d'autorité des localisations. Il est clair qu'une fois fait, ce choix est définitif sauf à refaire le chargement de toutes les données.
Le premier réseau Pancatalogue fut constitué avec autant d'institutions Dobis qu'il y avait de sections participantes et le pilotage des chargements était fait sur le numéro de la notice et sur la section de bibliothèque élevée au niveau d'une institution Dobis.

L'intégration des données bibliographiques

Passés les contrôles de chargement, les nouvelles notices bibliographiques à charger doivent être complètement réorganisées pour leur intégration. Comme pour les données de localisation, les transformations à faire sur les données bibliographiques sont fonction des caractéristiques du catalogage et du format d'origine, des contraintes internes de Dobis et du type d'affichage souhaité. Les traitements de conversion constituent une sorte de chemin critique pour obtenir le meilleurcompromis possible..
L'organisation des index de Dobis est le pivot de ces opérations. Au nombre de 9, ils constituent, notamment pour les auteurs, les titres, les titres de collection, les éditeurs, les sujets, autant de fichiers d'autorités intégrés. Ainsi, par exemple, chaque forme de nom d'auteur ne figure qu'une seule fois dans la base, quel que soit le nombre de notices auxquelles elle est associée et il en est de même pour toutes les données des autres index. Lors de l'intégration d'une nouvelle notice, chaque donnée qui la compose est comparée avec les données existantes de l'index correspondant. Si la donnée figure déjà, elle est réutilisée, pour la nouvelle notice. Seul est créé un nouveau lien entre cette donnée et le reste de l'enregistrement. Si, au contraire, la forme d'entrée n'existe pas, elle est créée dans l'index avec un premier lien vers le reste de l'enregistrement. C'est ce dispositif qui permet la fusion des données identiques dans les index et, en amont, les programmes de conversion réorganisent au mieux les données pour le plein effet de ce mécanisme. Dans la première version de Pancatalogue, le choix avait donc été fait de réutiliser le logiciel de conversion que Emory University (Atlanta, Georgie) avait développé pour le même usage, moyennant quelques adaptations pour le traitement des données bibliographiques propres à chaque bibliothèque dans les fichiers locaux de Dobis qui furent faites par l'auteur du programme.
La mise en place du dispositif informatique, l'installation de Dobis et du logiciel Emory modifié, l'établissement du réseau Dobis-Pancatalogue, l'adaptation des affichages se déroulèrent à partir de janvier 1988. L'initialisation de la base par des premiers chargements en vraie grandeur s'ensuivit et, en août 1989, on put croire, au vu d'une base expérimentale de 5 000 notices OCLC que la phase d'initialisation était achevée. Aussi la décision fut-elle alors prise de lancer à la fois les chargements opérationnels et les appels d'offres pour les nouveaux développements nécessaires à l'intégration des autres types de données. Mais la montée en charge de la base allait montrer qu'avant d'atteindre un état opérationnel, il serait nécessaire de faire de nouvelles mises au point. De plus, ce constat dut être fait par une nouvelle équipe.

Passés les contrôles de chargement, les nouvelles notices bibliographiques à charger doivent être complètement réorganisées pour leur intégration

Passage à la version opérationnelle

Le changement d'équipe

Cette fin d'année 1989 vit aussi le départ de l'équipe d'origine. Dans l'état d'avancement du projet, cette rupture fut un grave préjudice qui lui aurait sans aucun doute été fatal s'il n'avait bénéficié en contrepartie de deux facteurs favorables. La continuité de la volonté politique fut assurée par la nouvelle Direction de la programmation et du développement universitaire (DPDU) qui succédait à la DBMIST comme promoteur du projet et il faut souligner le rôle capital de l'agence IBM Lyon et plus particulièrement, l'aide essentielle apportée par l'ingénieur alors en charge des affaires du Sunist qui, de fait, assura la continuité technique du projet et permit la transmission du savoir à la nouvelle équipe.
Ce changement d'équipe s'accompagna aussi d'un changement de méthode. Désormais il y eut une forte implication du Sunist dans tous les aspects du projet avec la création, malgré toutes les difficultés dues à la séparation géographique, d'une véritable équipe de projet avec les compétences complémentaires d'informaticiens au Sunist et de bibliothécaires à la DPDU, qui prit en charge toutes les opérations.

Mise au point et mise en exploitation

Conformément au plan initial, les chargements opérationnels avaient commencé fin 1989. L'initialisation d'une base conçue pour comprendre des millions de données est, par nature, une opération lente et, de plus, l'analyse de la base résultat n'est valable qu'à partir d'un certain volume de données. En outre le déménagement des services du Ministère interrompit pendant plusieurs semaines tous les chargements encore pilotés depuis Paris à cette époque. Ce n'est donc qu'en mars 1990 qu'il fut possible de faire les premiers examens qui, pensait-on, seraient de pure forme, puisqu'avec les mêmes programmes Emory University avait créé

un catalogue de plusieurs centaines de milliers de notices OCLC .

En fait, l'étude de cette base constituée de 20 000 notices donna la mesure de la diversité des données reçues et de l'importance du travail à faire pour leur mise en cohérence. Les raisons pour lesquelles le logiciel satisfaisant aux besoins d'Emory University nécessita tant d'ajustements pour Pancatalogue sont de deux ordres. Tandis que l'université Emory intègre des données validées par une seule équipe de catalogueurs et les corrige en ligne sur son système local, Pancatalogue reçoit des données établies par plusieurs établissements qui cumulent dans leur quantité toutes les variantes autorisées, c'est-à-dire non contrôlées, par OCLC. A cela s'ajoute, comme nous le verrons plus loin à propos de la maintenance, la nécessité de faire toutes les corrections de données avant leur intégration, dans les programmes de conversion.

Pour ces corrections automatiques, la première difficulté est l'identification des variantes perturbatrices à partir de leurs effets dans la base. C'est dire qu'il s'agit de remonter des effets aux causes à travers toutes les étapes de la chaîne d'intégration. Pour cela, il est nécessaire de rentrer dans le détail intime des données et des programmes. Il est nécessaire aussi de travailler sur des quantités de données suffisantes pour pouvoir distinguer le caractère occasionnel ou répétitif des phénomènes et identifier les facteurs concomitants.

Une fois les diagnostics établis, la deuxième étape est de concevoir des mesures correctrices, c'est-à-dire des modifications et des compléments aux programmes de conversion, puis de les mettre en œuvre et de les tester. Après cela, il n'y a plus qu'à réitérer le processus, identifier de nouvelles variantes perturbatrices sur une base plus importante, ajuster ou compléter les traitements, recharger des données avec les nouveaux programmes et faire le diagnostic du nouveau chargement. Ceci jusqu'à l'obtention d'une base expérimentale assez propre et d'un volume suffisant pour qu'on puisse raisonnablement estimer que la qualité obtenue sera conservée lors de la montée en volume. La taille critique pour un diagnostic de valeur générale n'était pas connue à l'avance, elle s'avéra être de 200 000 notices.

Toutes les anomalies se révélèrent être dues à des différences entre les données réelles et leur définition théorique qui avait fondé l'analyse du logiciel. On peut citer par exemple la non-utilisation quasi systématique du sous-champ spécifique destiné à contenir les indications de volumes dans les titres de collections. Ceci générait autant d'entrées dans l'index qu'il y a de volumes dans la suite et on peut facilement imaginer l'effet produit dans l'index des titres de collection avec les volumes de la collection « Que-sais-je ? ». Les ponctuations parasites en fin de données comme les espaces, les virgules ou les points, qui génèrent autant de formes différentes et donc autant de doublons, nécessitèrent un nettoyage systématique. Le traitement des diacritiques identifiés comme tels, mais mal codés, est un autre exemple. Leur remplacement par un caractère unique (par exemple un dièse) pour éviter de rejeter la totalité d'une notice pour un seul caractère erroné fut une solution mise en place puis abandonnée, devant les perturbations induites à l'affichage par ce caractère exotique. La reconnaissance de toutes les variantes textuelles des mentions d'auteurs secondaires pour trouver leur équivalent codifié fut un autre problème.

Il faut imaginer aussi que toutes ces corrections portaient sur des matières encore peu connues de la nouvelle équipe et bien peu documentées, qu'il s'agisse des données et des analyses initiales, des programmes de conversion et de chargements, des caractéristiques de Dobis et, *a fortiori*, du rôle de chacun de ces facteurs dans les effets observés. A ces difficultés s'ajoutèrent aussi les surprises constamment renouvelées apportées par la disproportion entre les effets et les causes et, par conséquent, entre les causes et l'importance des traitements correcteurs à mettre en place.

Après 12 mois de mise au point (mars 1990 à mars 1991) et 3 mois de chargement intensif, enfin, en juin 1991, l'accès à un Pancatalogue de 200 000 notices OCLC, traitées selon des procédures entièrement automatiques, fut proposé en test à 3 bibliothèques (la bibliothèque interuniversitaire de sciences de Jussieu, la bibliothèque interuniversitaire de Lille et l'Institut national des langues et civilisations orientales). Après

Pancatalogue reçoit des données établies par plusieurs établissements qui cumulent dans leur quantité toutes les variantes autorisées

Pancatalogue
Calendrier de réalisation

Début	**1987**	Lancement du projet
Décembre	**1987**	Choix de DOBIS
Août	**1989**	1re base expérimentale 5 000 notices OCLC
Mars	**1991**	1ers chargements d'exploitation Version 1
Juin	**1991**	200 000 notices OCLC
Avril	**1992**	1ers chargements d'exploitation Version 2
Juin	**1992**	Etat de la base : 170 000 notices OCLC 70 000 notices BN 160 000 notices SIBIL
Novembre	**1993**	Etat de la base : 600 000 notices OCLC 150 000 notices BN 250 000 notices SIBIL

validation, le catalogue collectif fut ouvert en octobre à toutes les autres bibliothèques universitaires.
Pendant toute cette période supplémentaire de mise au point pour l'intégration opérationnelle des données OCLC, la phase suivante, déjà lancée depuis un an, avait elle même progressé.

Vers un catalogue multisource

Conformément au programme initial établi en août 1989, en même temps qu'avaient été lancés les premiers chargements, les appels d'offres avaient été initialisés pour la réalisation de la phase suivante. Du point de vue informatique, il s'agissait d'installer l'application sous la dernière version de Dobis (version 2) avec la reprise de tous les développements spécifiques réalisés sous la version antérieure, dont les logiciels de traitement des données OCLC. Du point de vue utilisateur, on peut dire que ce qui caractérise cette deuxième version est le changement d'alimentation avec l'intégration des trois types de données initialement prévues, BN-Opale, Sibil-France et OCLC. Cette nouvelle version allait aussi comprendre la refonte du système de pilotage des chargements.
Sur les 18 sociétés dont la candidature avaient été retenue, deux seulement firent une offre pour la réalisation des développements demandés, dont la compagnie IBM France avec laquelle un marché fut passé le 1er octobre 1990. Dans l'état de connaissance du moment, ce marché prévoyait l'ensemble des réalisations sur 12 mois, un avenant sera nécessaire pour adapter le contrat aux réalités.

Le nouveau réseau et le nouveau pilotage des chargements

Le changement d'alimentation augmentait considérablement le nombre des bibliothèques participantes et le problème délicat des localisations se posait donc à nouveau avec la montée en puissance de l'application. Il s'agissait dorénavant de pouvoir identifier au moins 300 sections de bibliothèques universitaires, sans compter, ultérieurement, d'autres unités documentaires des universités. Techniquement, il n'était pas possible de gérer un réseau Dobis/Pancatalogue de plusieurs centaines d'institutions, comme cela aurait été le cas en maintenant le choix initial de mettre chaque section de bibliothèques au niveau d'une institution Dobis. De plus, ce choix créait une distorsion entre l'organisation administrative des participants et celle du réseau technique, puisqu'en rendant les sections autonomes, il ne permettait plus une identification du fonds global de la bibliothèque.
Cependant, pour le prêt, il fallait pouvoir conserver l'affichage des localisations au niveau des sections. La solution mise en place fut de créer une base supplémentaire contenant l'ensemble des couples d'identifiants des données déjà chargées et constituant de ce fait un index de l'application. Cette base située en amont de Dobis est à la fois consultée et mise à jour à chaque chargement de notice. Cette solution rend plus complexe le dispositif d'intégration, mais, en répartissant les traitements des localisations, elle permet de lever une partie des contraintes contradictoires qui pèsent sur ces données.

Il s'agissait dorénavant de pouvoir identifier au moins 300 sections de bibliothèques universitaires

Une autre question importante touchant aux localisations est celle relative au traitement des cotes. Initialement, la présence des cotes s'était imposée par l'obligation faite alors au Pancatalogue de pouvoir restituer à chaque bibliothèque les données exactes qu'il avait reçues d'OCLC. Comme il s'était révélé impossible de mettre ces données à jour, seule la cote initiale figurait dans la première base. Depuis, il était devenu clair que la livraison des notices était du domaine des sources, mais l'intérêt des cotes demeurait pour le prêt entre bibliothèques, de même que demeu-

raient les difficultés pour les collecter et les gérer.
En effet, les bibliothèques disposant d'un système local ne seraient pas nécessairement enclines à faire systématiquement ajouts et changements de cotes dans les sources. De plus, cette difficulté de collecte supposée levée, la mise à jour des cotes par chargement restait problématique. Elle aurait nécessité que ces informations figurent dans la base index et constituent un des éléments clefs du dispositif devant alors accepter les données ne se différenciant que par la cote et les rejeter en cas de cotes identiques. Ceci n'était pas concevable compte tenu du caractère hétérogène de ces données non contrôlées ni contrôlables. Parce qu'au contraire, les cotes ou numéros d'inventaires sont des attributs locaux, que les demandes de prêt entre bibliothèques ne portent pas (sauf exception qui ne relève pas d'un catalogue collectif général) sur un exemplaire spécifique, parce qu'enfin, le fournisseur potentiel doit, en tout état de cause, interroger son système local pour vérifier la disponibilité du document demandé, il fut décidé que Pancatalogue ne recenserait que la présence des titres dans les institutions, indépendamment du détail des exemplaires.
Pendant que ces questions étaient instruites pour des choix qui détermineraient de façon définitive la nouvelle chaîne de traitement, l'implantation de la version 2 de Dobis sur l'ordinateur du Sunist, le paramétrage de ce logiciel, la reprise des modifications d'affichage et la mise en place de la base index progressaient. Progressaient aussi les travaux liés à l'intégration des données bibliographiques dans le nouveau contexte d'alimentation par plusieurs sources.

Les premiers résultats et la révision de la méthode

Pour cette nouvelle version, il avait été décidé de laisser coexister les notices de chaque type en juxtaposant les descriptions d'un même document provenant d'OCLC, de BN-Opale et de Sibil-France. Ceci devait conduire à un maximum de 3 notices pour un document et constituer des conditions encore convenables d'interrogation sous réserve que, dans les index-fichiers d'autorité, les données soient bien fusionnées.
On pouvait penser que l'intégration des données Sibil et BN-Opale dans Dobis nécessiterait globalement le même genre de traitements que ceux mis au point pour les données OCLC. On pensait aussi que le traitement de la ponctuation serait facilité dans le format Unimarc par une restitution entièrement automatique et qu'une bonne part des traitements de données OCLC seraient réutilisables pour les données Sibil qui pouvaient être livrées en USMarc. La méthode à utiliser semblait donc bien connue et le choix fut fait de réutiliser des programmes existants, développés par IBM pour d'autres applications sous Dobis, avec des adaptations dont nous pensions avoir la mesure avec l'expérience antérieure.

Les logiciels utilisés avaient été conçus pour intégrer des données différentes dans des bases différentes

Les premières intégrations de données Unimarc, leur comparaison avec les données équivalentes d'OCLC et de Sibil, révélèrent l'importance des travaux qui allaient être nécessaires pour obtenir la cohérence attendue. Se posa alors la question de la méthode pour y parvenir. En effet, les logiciels utilisés avaient été conçus pour intégrer des données différentes dans des bases différentes. Appliqués à l'alimentation d'une base unique, ils produisaient un catalogue tellement hétérogène qu'il était inutilisable.
Pour adapter ces logiciels, on pouvait faire des séries de mises au point par approximations successives, source après source, puis avec deux sources, puis, enfin, avec les trois ensembles de notices. On pouvait, au contraire, reprendre complètement les analyses de base et réviser de façon globale la conception des traitements. En tenant compte de la nécessité de disposer de documents de référence uniques qui servent aussi bien à l'élaboration des programmes, à leur mise au point, aux tests et ultérieurement à la maintenance et à l'administration de la base, le doute n'était pas permis. La reprise des analyses s'imposait, axée autant sur les caractéristiques des données d'origine, leurs points communs et leurs différences que sur celles du nouveau produit à construire.

Vers un nouveau catalogue

Le but des traitements était inchangé : il s'agissait de faire en sorte que les conditions d'accès, de lisibilité et de complétude de la base soient optimales. Les obstacles étaient constitués par le volume des données, plusieurs millions, et l'alimentation par trois types de notices. Les moyens passaient par l'obtention d'index de qualité présentant à la fois une bonne lisibilité et le minimum d'ambiguïtés et l'obtention de notices aussi uniformes que possible.
Qu'il s'agisse des notices considérées de façon globale ou qu'il s'agisse des éléments dans les index/fichiers d'autorité, la finalité des traitements est la même. Il s'agit de ramener chaque groupe de données à une forme commune choisie pour être la forme canonique de Pancatalogue. Il fallut ainsi définir la forme canonique des « auteurs personnes physiques », à laquelle seraient ramenées, autant que possible, toutes les données de ce type quelle que soit leur origine et il en fut de même pour les collectivités, les congrès, les titres des ouvrages, les titres de collections, les éditeurs, l'ISBN et autres codes chiffrés. De même, il fallut aussi définir la forme commune des nouvelles notices du nouveau catalogue. Il s'agissait bien en effet de définir un nouveau catalogue, même si les principes fondateurs émanaient autant des possibilités (et des impossibilités) de reconnaissance et de manipulations des données reçues que d'une réflexion théorique sur le contenu et l'organisation du nouveau système.
Cette nécessité d'établir, à la fois et

en même temps, les principes catalographiques fondamentaux et le détail des traitements de tous les éléments, cette obligation de trouver l'équilibre en marchant, qui ne constitue pas non plus une méthode classique en informatique est une caractéristique du projet. C'est aussi une cause importante des difficultés rencontrées, notamment pour prévoir les calendriers de travail et toutes les parties prenantes du projet ont dû s'adapter à ces conditions particulières. En effet, la connaissance des données concrètes sur lesquelles il fallait effectivement travailler ne pouvait qu'être progressive ce qui ne permettait pas aussi souvent qu'on l'aurait souhaité de dégager des conclusions définitives pouvant servir de fondation à la suite de la construction. A cela s'ajoutait l'obligation de mettre en place la totalité des traitements avant le début des chargements définitifs, parce que chaque modification des programmes d'entrées génère une nouvelle strate de données dans la base et perturbe ainsi sa cohérence, même si, par ailleurs, les nouvelles notices chargées prises individuellement en sont améliorées.

Il s'agit de ramener chaque groupe de données à une forme commune choisie pour être la forme canonique de Pancatalogue

Les choix fondamentaux et les traitements

On ne peut faire ici l'exposé complet des questions et des choix, pas plus que celui du détail des traitements. La liste en serait fastidieuse, mais, de plus, leur compréhension nécessiterait de connaître aussi le détail des contraintes à prendre en compte comme les caractéristiques de Dobis, les formats et règles d'origine et celles liées au souci de dénaturer le moins possible les données reçues. Les exemples ci-dessous illustrent la problématique et les méthodes utilisées.

Une question fondamentale porte sur le choix des données à conserver dans le nouveau catalogue. Elle se pose aussi bien pour des zones entières de la notice que pour des données élémentaires dans les index.

Dans les noms personnels par exemple, les données d'origine sont plus ou moins riches d'indications complémentaires comme les dates de naissance, de décès, les qualificatifs de profession ou d'état. Choisir, comme forme canonique, le plus petit dénominateur commun c'est-à-dire – nom et prénom – aurait permis la meilleure fusion de toutes les données dans leur index et conduit à une excellente lisibilité, mais cela aurait conduit aussi à des fusions incorrectes sur des homonymes que rien n'aurait alors plus distingués les uns des autres. En revanche, conserver tous les éléments aurait généré trop d'entrées, de formes voisines, pour un même auteur et perturbé l'interrogation en proportion. La forme canonique retenue fut donc un moyen terme composé des nom, prénom et date de naissance et les variantes à ce modèle ne sont donc dans Pancatalogue, aux erreurs près dans le codage d'origine, que des variantes par défaut.

Le problème est identique pour ce qui concerne la sélection de zones entières comme par exemple celles des titres. L'étude des données reçues mit en évidence que les titres dits uniformes ne le sont que par rapport à un système de référence donné, qu'un titre original peut en cacher un autre, notamment dans le cas de traductions en cascade, qu'un titre de regroupement n'a de sens que pour un catalogue donné. Il apparut en outre que la présence de ces entrées était loin d'être systématique. Dans ces conditions, on pouvait s'interroger sur l'intérêt d'un regroupement partiel d'un sous-ensemble indéterminé de notices. A quoi peut servir, par exemple, la présence occasionnelle d'entrées normalisées de type *Bible* ? Il fut alors décidé de ne conserver dans Pancatalogue que les zones de titres strictement relatives au document décrit.

Les données à intégrer font aussi l'objet de traitements qui passent toujours en premier lieu par le nettoyage de caractères parasites dont, contrairement à ce qu'on avait pu imaginer au départ, on ne peut s'affranchir pour aucune d'elles. Ils peuvent aussi comprendre une restructuration : ainsi, les subtilités qu'implique la distinction entre sous-titre, complément du titre et reste du titre conduisent à des choix différents d'une notice à l'autre, *a fortiori* d'une source à l'autre. Pour permettre une bonne fusion des données dans les index, mais surtout pour harmoniser les possibilités d'interrogation à partir des mots des titres, ces éléments sont le plus souvent regroupés. Pour d'autres raisons, mais pour le même objectif, il faut réorganiser systématiquement les sous-champs des congrès. Parce que de telles décisions peuvent mettre en cause la réversibilité de la conversion, leur usage en a été très limité.

Les données à intégrer font l'objet de traitements qui passent toujours en premier lieu par le nettoyage de caractères parasites

Il est parfois nécessaire de reconstituer des données manquantes. C'est le cas par exemple pour un code indiquant si le nom personnel d'auteur est simple ou composé. Nécessaire à Dobis, mais absent des données Sibil et BN-Opale, il est reconstitué à partir des données elles-mêmes où, après nettoyage, on recherche la présence d'espaces ou de tirets dans la chaîne de caractères qui compose le nom. Constatant ensuite que cette reconstitution est plus fiable que les codages manuels, la réutilisation du code correspondant des données OCLC fut abandonnée et la méthode est mainte-

nant généralisée. Un autre exemple est le code de langue. Les notices d'origine ne contiennent qu'une seule indication de langue s'appliquant à l'ensemble du document, alors que, dans Dobis, chaque donnée auteur, titre et matière doit être caractérisée spécifiquement pour l'organisation des permutations et des accès sur les mots non vides des index correspondants. La restitution d'un code fonctionnel était donc indispensable, elle fut réalisée suite à une analyse des données et des mécanismes de Dobis, mais plus encore grâce à de très nombreux essais.

Il convient aussi de faire une mention spéciale des données matières qui, à tous égards, constituent une catégorie particulière. Ni purement descriptives ni purement de regroupement, ces données présentent de nombreuses difficultés de traitement dues à la faiblesse de leur normalisation. Ce défaut de normalisation joue à la fois sur le contenu des données, sur la définition des formats et aussi sur leur organisation dans les systèmes informatiques avec des effets qui se cumulent. C'est en raison de ces particularités que, faute de temps, les traitements sur les données matières sont limités à un objectif de lisibilité et la comparaison des noms personnels dans les index auteurs et matières illustre la nature et la nécessité des traitements sur ces données (cf. encadré ci-dessous).

La maintenance catalographique

Le rôle stratégique des données de localisation et les transformations faites sur les données bibliographiques posent de façon particulière la question de la maintenance catalographique d'une base comme Pancatalogue.

Le premier point à souligner est le suivant. Qu'il s'agisse de remplacement, de correction ou de suppression, les interventions s'exerçant sur les données après leur chargement modifient l'ordonnancement de la base et sont donc des actions à risque. Ainsi, par exemple, toute modification de localisation touche au système de pilotage des chargements, de même, une suppression de notice. En conséquence, les procédures correspondantes sont en général complexes, elles nécessitent beaucoup de précautions et certaines même sont à proscrire.

Ce point établi, se pose la question des modalités. Trois modes d'action sont envisageables : des chargements de données nouvelles venant remplacer des données existantes, des interventions faites en ligne par des catalogueurs correcteurs et l'activation de programmes spécifiques.

La première voie consiste à remplacer systématiquement ou selon des critères définis des données existantes par des données nouvellement livrées. L'intérêt d'une telle opération suppose que le gain en qualité soit certain et assez important pour justifier le surcoût des chargements supplémentaires. Or, en raison des transformations faites sur les données reçues pour les harmoniser avec les notices autres, il n'est pas garanti que toutes les corrections faites dans les sources soient utiles et adaptées au nouveau catalogue. A cela, il faut ajouter que la qualité initiale des notices est élevée et qu'il est difficile de repérer de façon simple sur quoi porte la différence entre deux états d'une même notice. Pour ces raisons, c'est la première notice chargée qui est conservée aujourd'hui dans Pancatalogue.

Les corrections faites en ligne par des catalogueurs correcteurs pourraient constituer une méthode efficace dans une situation stationnaire ou quasi stationnaire. Mais compte tenu du volume des nouvelles entrées, elle organiserait de fait une course poursuite entre les correcteurs et les chargements de données, chaque chargement mettant en cause le gain de cohérence obtenu par les correcteurs. Il faut aussi noter que la croissance du catalogue nécessiterait un nombre de correcteurs augmentant en proportion. Les interventions humaines seront donc réservées à quelques cas résiduels après que toutes les corrections possibles par programmes auront été réalisées.

Ce troisième mode de maintenance catalographique – la correction de données par activation de programmes – est aujourd'hui un chantier ouvert, dont les premiers résultats sont déjà appréciables sur la base actuelle. Ainsi, par exemple, tous les

Index matière : Néel (Louis), 6 formes pour 6 occurrences

Recherche
Mots Matières

1	Basic	needs. -y	
2	Cheney, Frances	Neel - (1906-) Correspondance. ram	
3	David-	Neel - Alexandra 1868-1969. ram	
4		Néel - Louis 1904-. ram	1
5	---Cheney, Frances	Neel, - 1906- Correspondance ----------	
6	David-	Neel, Alexandra - (1868-1969). ram	
7	David-	Neel, Alexandra - 1868-1969	
8	David-	Néel, Alexandra - (1868-1969). ram	
9	David-	Neel, Alexandra, - 1868-1969	
10	----------	Néel, Louis - (1904-) - biographies. ram. ----------	1
11		Néel, Louis - (1904-....). ram	1
12		Néel, Louis - 1904-	1
13		Néel, Louis. ram	1
14		Néel, Louis, - 1904-. ram	1

Entrez le numéro ou le code choisi

t terme f suite
i index b retour u créer lot
d détail e fin

codes sur la forme des noms d'auteur, nom simple ou nom composé, ont été corrigés de cette façon et les permutations correspondantes complètement rétablies. D'autres corrections de ce type sont à l'étude qui constitueront en outre une expérience précieuse pour des opérations de plus grande ampleur.

L'application aujourd'hui

Aujourd'hui, Pancatalogue compte plus d'un million et demi de documents localisés dans quelque 80 bibliothèques universitaires ou sections. Il est régulièrement alimenté par la production catalographique courante que ces bibliothèques réalisent dans les trois sources de catalogage reconnues OCLC, BN-Opale et SIBIL-France. A ces nouveautés s'ajoute la production du programme national de rétroconversion des fichiers des plus grandes bibliothèques et, pour les prochaines années, l'accroissement annuel prévu est de 700 000 nouveaux documents.
De façon succincte, on peut dire que Pancatalogue sous Dobis, c'est 250 fichiers, 1 200 modules PL/1 de programmes internes et 250 modules PL/1 pour les programmes de conversion des données. L'application fonctionne avec deux bases, l'une organisée pour la consultation occupe environ 3 giga-octets, l'autre organisée pour les chargements occupe le double. Les chargements sont quotidiens et, périodiquement, la base de chargement est basculée en consultation.

L'accès à Pancatalogue est possible avec toutes sortes de terminaux via Transpac et Renater et aussi, depuis l'étranger, par Internet

L'accès à Pancatalogue est possible avec toutes sortes de terminaux *via* les réseaux français Transpac et Renater et aussi, depuis l'étranger, par Internet. L'interrogation de la base est possible en mode professionnel (Dobis) et aussi en mode public (Libis) plus particulièrement développé pour les interrogations à partir d'un minitel professionnel (80 colonnes).
Les services rendus par ce catalogue collectif, notamment pour le prêt entre bibliothèques, ne peuvent pas encore être connus avec précision. Ils sont en effet subordonnés à l'obtention d'une taille suffisante de la base, qui sera très prochainement atteinte, avec, entre autres, les opérations de rétroconversion. On constate cependant une augmentation régulière de son utilisation tant par les abonnés que par les usagers plus occasionnels sur le 36-17 code Panca.
Pour ce qui concerne les bibliothèques participantes, on peut regretter que l'obligation qui leur est faite de travailler dans une source de catalogage partagé limite leur nombre. Mais, comme on l'a vu avec la description des traitements nécessaires à l'intégration des données, il est difficilement envisageable de pouvoir obtenir la même qualité de catalogue et les mêmes facilités d'interrogation à partir de données moins rigoureusement normalisées.

L'avenir

Trois registres de considérations sont à prendre en compte pour l'évolution de Pancatalogue. Certaines améliorations se situent dans la logique même de l'application et permettraient que soient mieux satisfaits les objectifs initiaux. Parmi celles-ci figurent la fusion des données bibliographiques pour obtenir une seule notice par document avec le regroupement de l'ensemble des localisations et, tout aussi indispensable, la mise en place de procédures pratiques et sécurisées pour répercuter les changements de localisations des collections dans les bibliothèques participantes.
De tels objectifs comme les voies et les moyens pour les réaliser sont à instruire dans le cadre du schéma directeur informatique des réseaux de bibliothèques universitaires. On peut dire que, dans tous les cas envisagés, les développements ci-dessus constituent un préalable indispensable à la réutilisation des données sous un nouveau système informatique.
Le troisième registre de réflexion dépasse le cadre de l'enseignement supérieur et touche au Catalogue collectif de France. Les liens entre Pancatalogue et ce futur service ont été définis par une décision interministérielle (septembre 1991) qui établit Pancatalogue comme une composante du Catalogue collectif de France. C'est d'ailleurs en application de cette déci-

Pancatalogue

ACCES

avec contrat : (à demander au SUNIST)
- Minitel 1B (Minitel professionnel, affichage 80 colonnes) :
3613 code d'accès **SUN1**, service **PANCA**
- Mocro-ordinateur + carte de télécommunication en émulation VT100 :
numéro d'accès à Transpac **36062424**
numéro d'appel du Pancatalogue **134022271494**
- Configuration spéciale pour utilisation du réseau privé X25

sans contrat :
Minitel 1B (Minitel professionnel, affichage 80 colonnes) : **3617** code **PANCA**

CORRESPONDANTS :

Ministère Enseignement supérieur et Recherche :
Anne-Marie Motais de Narbonne, Tél : (1)-40-65-60-89
Mireille Penichon, Tél : (1)-40-65-62-79

SUNIST :
Olivier Serre, Tél : (16)-67-14-86-05 - (16)-67-14-14-14

sion que s'inscrit la convention passée en juin 1992 entre le Ministère et la Bibliothèque de France sur une programmation commune des opérations nationales de rétroconversion des catalogues des universités destinées à alimenter l'un et l'autre de ces catalogues nationaux.

Conclusion

Au-delà du partage des ressources, la coopération entre les bibliothèques est destinée à satisfaire les besoins de la recherche en rendant accessible l'ensemble des fonds documentaires des universités à l'ensemble des chercheurs. *In fine,* ceci est concrétisé par la fourniture des documents, mais cela passe aussi par des moyens performants et sûrs pour identifier et localiser ces documents. Ces moyens doivent notamment permettre, par une seule démarche de l'usager, aussi bien le regroupement de toutes les données répondant au critère choisi que le pointage direct sur un document déterminé.

Aujourd'hui, la méthode utilisée pour atteindre cet objectif est de cumuler les données dans un catalogue collectif et nous avons vu que les difficultés de l'entreprise tiennent essentiellement à la nécessité de concevoir des traitements entièrement automatiques pour des données hétérogènes.

Demain, peut-être, le même objectif pourra être satisfait non plus par le traitement systématique et *a priori* des données, mais par celui au coup par coup des requêtes qui seraient alors successivement transférées sur chaque catalogue de bibliothèque et par celui des réponses partielles ainsi obtenues. Le scénario met en lumière que les difficultés rencontrées pour l'établissement de Pancatalogue tiennent moins à la méthode qu'à l'objectif poursuivi. Qu'il s'agisse aujourd'hui d'organiser des données dans une nouvelle base ou qu'il s'agisse demain d'adapter les requêtes à la variété des catalogues interrogés et de synthétiser les réponses partielles pour construire une réponse globale satisfaisant les critères indiqués, la problématique est la même.

En effet, si l'informatique est l'outil indispensable à l'ampleur de telles entreprises, si elle seule permet une véritable accessibilité des données, elle laisse cependant sans solution les questions fondamentales de cohérences intellectuelles et formelles et, plus encore, en amplifiant tous les effets, elle en souligne toutes les faiblesses. Ainsi, pour profiter des performances croissantes des techniques informatiques et de télécommunication, outre les efforts corollaires en matière de normalisation des données, il apparaît indispensable que soit poursuivie la réflexion classique fondamentale sur les catalogues, leurs rôles et leur utilisation.

Novembre 1993

BIBLIOGRAPHIE

1. Pennel (Patrice), Lupovivi (Catherine) , Denis (Anne-Marie), « Le Plan catalogue », *Bulletin des bibliothèques de France,* Paris, t. 32, n° 2, 1987, p. 118-132.

2. *Les Bibliothèques universitaires,* sous la direction de Daniel RENOULT, Paris, Cercle de la librairie, 1994, (Collection Bibliothèques). A paraître.