Ar(abes)ques

JANVIER 1997 - JUIN 2018

DOSSIER

«Où m'ont-ils mis mon dollar d ?»

Sept mois de reprise des données

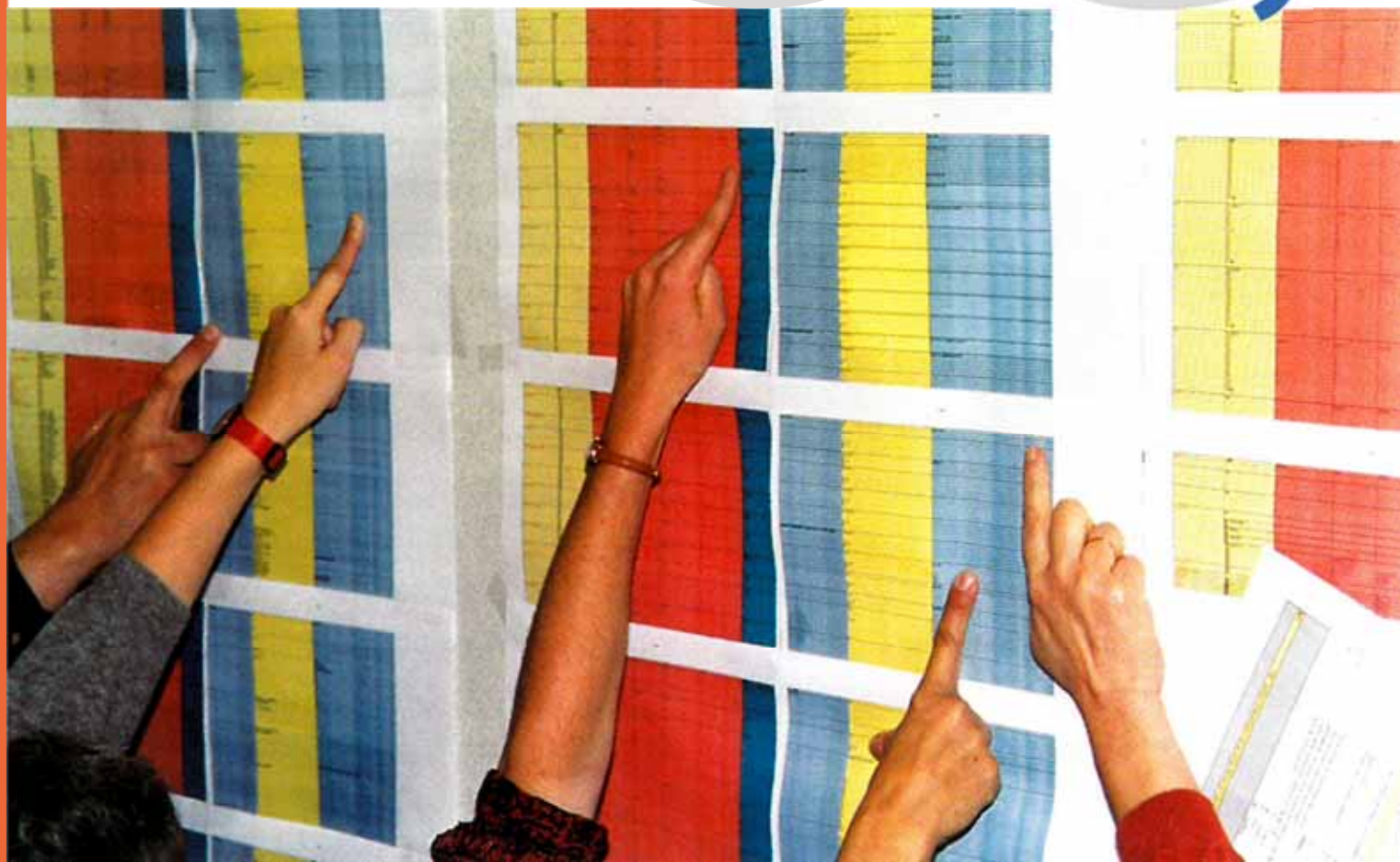
PLEINS FEUX SUR • Matière en fusion

ACTUALITÉS • Eplucher, émincer, cuire...

Numéro spécial

Tiré à part

abes
agence bibliographique de l'enseignement supérieur



«Où m'ont-ils mis

Sept mois de reprise des données*

Mai ⇨ novembre 1998 ⇨ **mars 1999**
Création de la base test : du 12 décembre 1998 au **23 février 1999**
Création de la base réelle : du 8 janvier 1999 au 13 mars 1999

En mai, cueillette d'exemples

Au mois de mai, la reprise cueille ses jeux d'essai. Le calendrier du projet nous donne la date limite du 4 juin 1998 pour envoyer à *Pica* 1 000 notices pour chacune des *7 bases à convertir.*

Les notices sont choisies avec soin : il faut qu'elles soient représentatives de chaque base. Tous les champs possibles doivent être présents au moins une fois. Différents cas de figure ont déjà été détectés, comme l'horreur des multivolumes, les divergences de catalogage entre les bases, les notices de rétroconversions, les différents supports, les notices anciennes, les langues gourmandes en diacritiques ou translittérées... Chaque membre de l'équipe est spécialisé dans l'une des bases, mais nous sommes 4 (2,9 équivalent-temps-plein) pour les 7 bases. Qui s'occupe du ***CCNPS***, qui d'***OCLC***, qui de ***TELETHESES***, et qui de ***BN-OPALE***, de ***SIBIL***, et des autorités (**Autorités de la *BNF*** et ***FMESH***).

Ces notices doivent aussi servir aux tests de dédoublonnage.

Nous cherchons donc la notice du même document dans les bases bibliographiques " concurrentes ".

Le pire qui puisse arriver est d'avoir un quadruplon (*BN-OPALE, OCLC, SIBIL, TELETHESES*).

A l'aide d'outils très sophistiqués - un bon vieux classeur papier pour les impressions de notices et *Excel* pour consigner leurs numéros - chacun cherche dans " sa " base des notices choisies par l'un de nous. Chaque exemple est vérifié, les doublons, triplons... sont agrafés en liasse, le tout dûment répertorié par numéro d'ordre. Pour plus d'amusement, nous avons glissé dans le colis quelques faux doublons.

Entre deux promenades sur les bases, agrémentées par les ralentissements d'internet ou l'absence de certaines données codées sur le cédérom pour *OCLC*, et par la différence de format pour *SIBIL* (consultée en *SIBIL-Marc* et exportée en *Unimarc*), il faut chercher les notices autorités correspondant aux notices bibliographiques. Car le jeu d'essai est aussi destiné à tester le paramétrage des liens notices bibliographiques / notices autorités. Ce travail fastidieux est le bizutage original d'une nouvelle collègue.

Nous voici donc munis de listes de numéros de notices, que nous envoyons au gestionnaire informatique de chacune des bases : c'est le coeur du test, des exemples que nous reconnaîtrons à chaque passage. Le chiffre de 1000 est atteint grâce au gestionnaire de chaque base, qui complète le lot.

.../...

Les 7 bases à convertir

1. *AUTORITES DE LA BNF*
   Auteurs, titres uniformes, *RAMEAU*
2. *BN-OPALE*
   Base bibliographique de la Bibliothèque nationale de France
3. *CCNPS*
   Catalogue collectif national des publications en série
4. *FMESH*
   French Medical Subject Headings* - autorités matière de la *USNLM
   (*United States National Library Medical*)
5. *OCLC*
   ***Online Computer Library Center* - Notices bibliographiques des bibliothèques du réseau AUROC (Association des utilisateurs du réseau *OCLC*)**
6. *SIBIL*
   Notices bibliographiques du réseau *SIBIL* France
7. *TELETHESES*
   Signalement des thèses soutenues en France

mon dollar d ?»

En guise de récréation, nous préparons la logistique des tests

Pica nous informe que les fichiers de notices converties ne peuvent sortir que sur une imprimante 138 colonnes. Le *CNUSC* qui en possède une a beau être tout près, nous ne voulons pas prendre une voiture pour aller imprimer une notice, ni contrôler sur papier 1 000 x 7 notices en 3 formats (imaginez le nombre d'arbres en cause).

Pourquoi 3 formats ?

Le premier est le format source, le deuxième le format interne Pica+, le troisième l'Unimarc. Nous décidons de demander uniquement les fichiers et de ne travailler que sur écran.

Il nous faut vraiment nous familiariser avec le format interne Pica+, que plus personne ne verra ensuite. Ce format présente des aspects déroutants : étiquettes à 4 caractères, vocabulaire inattendu...

Tout se passe bien, Pica a les échantillons à la date fixée. En route pour Babel-Marc.

En juin, récolte de notices

En juin, Pica nous envoie les premières versions des notices. Effroi : les tables de conversion accompagnant les notices sont les algorithmes du convertisseur, dont on ne comprend que les " if " et autres " else ", et les notices n'ont pas la fameuse 3e colonne en Unimarc. Réclamations, deuxième envoi avec des tables lisibles par un être humain, ouf ! Oui, mais. Comment ouvrir un fichier de 5 à 7 *Mo*. comportant 3 colonnes tabulées ? C'est *Excel* qui nous sauve, mais il faut décompresser le fichier, le tronçonner par lots d'environ 250 notices, pour enfin découvrir chaque notice en 3 formats. Un " léger " inconvénient subsiste : pour retrouver une notice, il faut chercher dans chacun des 5 fichiers.

Les notices *BN-OPALE* arrivent les premières le 11, suivies en rafale par *SIBIL* le 12, *OCLC* et BN-Autorités le 15, *CCNPS* et *OCLC* zones locales le 26, FMeSH le 30.

Les bureaux sont couverts par des couches de documents : les tables de conversion, le format d'origine de chaque base, le format Unimarc pour tous, les impressions des notices coeur-du-test. Les échanges avec Pica décollent, chaque jour amène son lot de questions/réponses. Le calendrier établi par le " comité reprise " (ABES/Pica) prévoit une semaine d'examen par l'ABES pour chaque conversion. Cela se traduit par une bonne dizaine d'heures par jour à examiner les jeux d'essai.

Et *TELETHESES* ? Sa conversion est repoussée pour cause de transmission : un esprit malin subtilise les marques de fin de notice entre l'ABES et Pica.

Nous apprenons par expérience à trier les notices par champ Pica+, par champ base source, par champ Unimarc : tous les recoupements sont possibles, et le "copier/coller/rechercher" devient notre drogue.

Nous communiquons par des grommellements: " mais où m'ont-ils mis le $d? ", " mais pourquoi ils ont fait ça comme ça ? ", " oui, mais en USmarc c'est comme ça"...

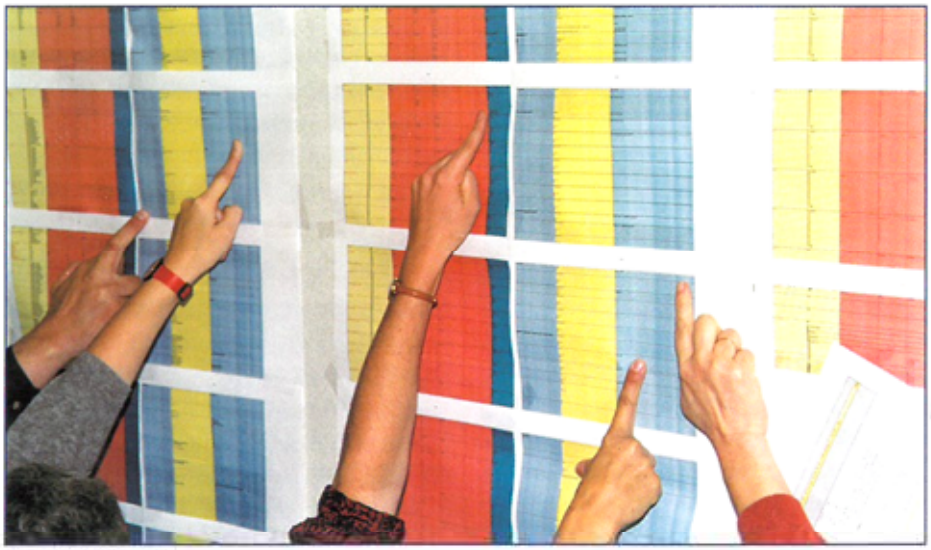
Chaque conversion donne lieu à un document rédigé par l'ABES pour Pica où chaque zone et sous-zone est commentée.

En juillet, nous sommes mûrs

En juillet, l'équipe ABES de la reprise ne peut pas partir en vacances, pour cause de conversions. L'équipe de Pica non plus, d'ailleurs. L'équipe de Pica lit nos commentaires, et les prend en compte.

Le 8, nous sommes excités : c'est le jour de la première visioconférence entre les deux équipes reprise. Nous ne nous connaissons que par messagerie électronique. Nous ne voyons que le sommet de la tête de l'un de nos interlocuteurs, car notre propre image est incrustée sur son visage. Mais c'est réconfortant de voir que nos interlocuteurs existent vraiment, ce dont nous doutons les jours de grande paranoïa.

Pica nous envoie une deuxième mouture des notices, avec les tables de conversion correspondantes. Un nouveau tir groupé : le 9 *BN-OPALE*, *SIBIL*, *CCNPS*, qui bénéficient du même convertisseur ; le 20 *OCLC*, le 22 FMeSH, le 24 les autorités *BN-OPALE*. L'entraînement a été bon, nous ouvrons les fichiers du premier coup.



De gauche à droite
«Christian Chabillon, Mireille Teissèdre, Cécile Laube, Geneviève Boyer, Elisabeth Coulouma,» devant leur mur de zones, sous-zones, valeurs de zones codées

Sur la «une», ***graphique de M. Teissèdre, d'après les statistiques de la base SIBIL au 4-11-98***

C'est le suspense : tout sera-t-il parfait? Eh bien non. Nous ne comprenons toujours pas comment Pica convertit les données codées, nous trouvons bizarre le traitement des dates, nous ne trouvons pas tous les indicateurs, des sous-zones manquent toujours.
Il reste à améliorer deux points importants : la conversion des zones *RAMEAU*, avec la possibilité de faire plusieurs liens vers des autorités dans la même zone 6XX, et la convergence des conversions USmarc/Unimarc. Mais il y a une réelle progression : la première conversion (FMeSH) est validée le 29.

LES 3 ÉTAPES DE LA " REPRISE "
1. conversions
2. chargements/dédoublonnage
3. liens entre les notices

En août, Pica travaille seul

Début août, l'équipage reprise est réduit à un matelot de quart, qui assure.
Le ***RBCCN*** est validé après quelques échanges de messages le 12.
Pour le reste, nous avons besoin d'oublier l'existence même des formats.
La première conversion de *TELETHESES* est livrée le 31.

DES MOTS POUR LE DIRE
CNUSC
Centre national universitaire sud de calcul
RAMEAU
Répertoire d'autorité matière encyclopédique et alphabétique unifié
RBCCN
Répertoire des bibliothèques membres du réseau du catalogue collectif national des publications en série
SU
Système universitaire de documentation

En septembre, égrappage de données codées

En septembre, l'équipe reprise s'agrandit. Nous sommes 5 maintenant, ce qui représente 3,9 équivalents temps plein. Le travail est mieux réparti : la nouvelle venue va s'occuper des autorités (FMeSH et *BN-OPALE)* et du livre ancien sur toutes les bases.
La troisième mouture des notices n'arrive toujours pas.
Par contre la deuxième version des notices de *TELETHESES* nous parvient le 25.
Entre-temps, nous continuons un travail peu alléchant, commencé en juillet/août. Pour contrôler la convergence de toutes les conversions des notices bibliographiques, nous avons décidé de faire un tableau synoptique, comportant toutes les zones et sous-zones (ou valeurs des zones codées) de l'Unimarc, et en regard les zones, sous-zones et codes de chacune des bases à reprendre, ainsi que le résultat en Pica+. Nous pensions qu'il fallait une bonne semaine, plus d'un mois est nécessaire. Le plus long est bien sûr de faire concorder les zones codées en Unimarc et USmarc : on voit bien que les bibliothécaires sont pervers !
Fin septembre, la première version de ce tableau multicolore, affiché sur 2 m², intrigue tous les collègues non concernés. Son petit frère, le tableau de concordance des notices autorités (format Intermarc, Pica+ autorités, Unimarc de catalogage autorités) a demandé à sa créatrice autant de doigté et de patience qu'un *Rubick's cube*.

Le projet du format de catalogage Unimarc a été envoyé par Pica : une bonne nouvelle, Pica gérera désormais presque tous les codes en Unimarc ; une autre bonne nouvelle, le format Pica+ a été remodelé pour mieux coller à l'Unimarc. Mais, pour nous, la pilule est amère : il faut vérifier à nouveau toutes les conversions dans ce Pica+ remanié. Et si une sous-zone restait coincée quelque part dans un convertisseur ?

En octobre, les formats fermentent

Chez Pica, notre principal interlocuteur, privé de vacances tout l'été, est absent jusqu'au 20 du mois. Son retour va-t-il coïncider avec une nouvelle avalanche de notices ? Non, car il peaufine ses conversions, là-bas à Leiden. Nous devenons incollables sur les zones codées, notre grand tableau comporte maintenant 1200 lignes et 11 colonnes. Et les notices nous manquent toujours.
Nous participons à la réalisation du " dossier de test d'intégration " de la base du *SU* par le maître d'œuvre : Pica devra vérifier que tout fonctionne comme nous le prévoyons. Seront testés les programmes spécialement développés (lien d'une zone de notice bibliographique à plus d'une autorité pour *RAMEAU*), et les paramétrages effectués (pour le dédoublonnage de chaque base).

En novembre, chauds les marrons

Le 2 novembre est pour les " unimartiens " la fête des notices : enfin les 3e versions arrivent. Les autres collègues sont jaloux. Le " comité reprise " des 3 et 5 novembre re-planifie la suite des opérations. Les notices de tests doivent arriver le 17. Au moment où vous lirez ce numéro d'*Arabesques*, nous en saurons plus sur la suite.

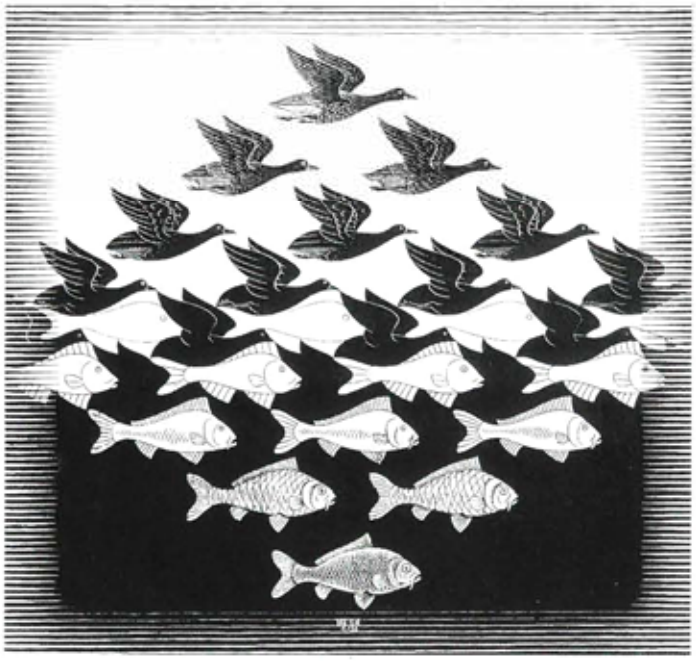
Et ensuite ?

Nous sommes en train de terminer le marathon par un sprint, ce qui exige un entraînement différent. Les opérations de création de la base test vont s'enchaîner jusqu'à fin février. Décembre nous trouvera encore penchés sur les validations des conversions des notices tests (fin le 12 / 12), le dédoublonnage suivra, la création des liens entre notices bibliographiques et autorités couronnera l'opération. Nous allons travailler en " flux tendu ". Chaque fois qu'un algorithme de dédoublonnage sera validé, la base réelle correspondante sera chargée. Cette opération de chargement commencera fin janvier. L'éloignement et le problème de la langue (l'équipe reprise chez Pica ne parle pas français) rendent les travaux plus longs et lourds. Pourtant, l'ennui est bien la seule chose qui ne nous guette pas.

Mireille Teissèdre
Mireille.Teissedre@abes.fr

* Sept mois de reprise des données
L'ABES, maître d'ouvrage, doit fournir à Pica, maître d'œuvre, les données constitutives du système universitaire de documentation : données existant dans *plusieurs bases de données* , " reprises " pour élaborer *un seul catalogue*.

Une matière en fusion



Le ciel et la mer I - **M. C. Escher,** ***l'œuvre graphique : introduction et commentaire du graveur,*** **Benedikt Taschen**
© 1990 Benedikt Taschen Verlag Berlin GmbH - D-1000 Berlin 10
© 1989 Reproduction de M. C. Escher : Cordon Art - Baarn - Holland

Suite de l'épisode précédent. Fin novembre 1998, les conversions des bases reprises sont toujours en cours. Malgré notre impression de piétinement, le projet avance : nous voyons que seuls quelques détails restent à régler pour les conversions bibliographiques. Entre-temps, les formats de catalogage et le dédoublonnage ont requis toute notre attention.

Des formats sans déformations

Formats de catalogage : notices bibliographiques, données locales et d'exemplaires, notices autorités, notices de centres de ressources.

Formats d'exportation : les mêmes, en ajoutant l'USMARC bibliographique et autorités.

Quelques zones nous donnent du fil à retordre depuis le début : la 240 (*auteur / titre*) des notices autorités et sa commère la 604 (zone correspondante en notice bibliographique) font l'objet de discussions internes, puis avec Pica. Nous marchandons une dernière modification en avril.

Les données codées, elles, sont les héroïnes d'une longue saga. Finalement converties et gérées en *UNIMARC* sans passer par l'étape du format interne, elles doivent être affichées dans des fenêtres qui rendent la saisie et la modification plus confortable. Mais disparaît au passage la *langue de catalogage* (obligatoire en *UNIMARC*). Puisque c'est toujours le français, sommes-nous tentés de dire, elle sera générée par défaut à l'exportation. Oui mais, les notices importées d'OCLC - Online Computer Library Center -, par exemple ? Certaines sont cataloguées en anglais, et la notion de valeur par défaut n'est finalement pas pertinente. La langue de catalogage est incorporée au format de catalogage au cours d'une réunion de « résolution de remarques » avec Pica.

La fonction de catalogage permet de valider les notices saisies, et pour cela un contrôle s'opère dans un « serveur de validation ». Bien sûr, cela ne se fait pas selon la tête du catalogueur, mais par des tables que Pica construit et nous demande d'approuver.

Ces notices vont comporter des liens (avec les autorités, ou entre elles). Cela aussi est contrôlé par des tables soumises à notre sagacité. Il faut en complément définir ce que verra l'utilisateur à la place du numéro de la notice liée (quelles zones / sous-zones de cette notice) dans chacun des formats.

Tout cela réglé, encore faut-il retrouver les notices dans la base. Les index à créer sont un monde à eux tout seuls ! Et selon les possibilités du logiciel, par exemple le fonctionnement des index conditionnels, nous pouvons remettre en cause les conversions elles-mêmes : chez Pica, Monsieur conversions c'est Sisyphe.

Détecter les doublons

Les utilisateurs du SU doivent pouvoir profiter d'une base déjà riche du travail antérieur ; mais riche ne veut pas dire encombrée. Le dédoublonnage est donc un élément majeur de ce projet.

Comprendre les algorithmes. La documentation standard sur le dédoublonnage, reçue début janvier 99 nous paraît difficile à comprendre. Passe encore pour le format interne, cela ne nous arrête plus. Mais il est question de statut, de poids, d'élément. Et dans les usages de nos interlocuteurs, « title » peut désigner la notice bibliographique ou le titre du document décrit...

Nous demandons une réunion « en chair et en os » à Pica, car il nous faut un vrai dialogue. Cette réunion a lieu les 28 et 29 janvier à Leiden, où se rendent deux d'entre nous. Nous pouvons ainsi voir le logiciel tourner en direct, faire des simulations, discuter de la trentaine de questions envoyées avant le voyage, et avoir immédiatement chaque réponse. C'est une réunion très fructueuse.

Le logiciel standard doit être adapté à notre format, à nos règles de catalogage, à notre langue. De retour à l'ABES, nous attendons les modifications. Deux tables sont utilisées, l'une pour les publications en série et l'autre pour les monographies.

Avoir la moyenne. Le premier ensemble de données à être versé dans le SU est extrait de ***BN-Opale***. Ces notices servent de référence. Une de leurs qualités est d'avoir le numéro de la notice liée dans la zone de la notice bibliographique : immense avantage d'un lien plus sûr que sur une chaîne de caractères.

Chacune des notices ensuite versée dans la base sera comparée à toutes les notices résidantes. Les index sont peu nombreux, ils ne servent qu'à identifier les « candidats » doublons : *ISSN et ISBN, clé titre.* Les candidats doublons sont les notices qui ont une ressemblance suffisante pour être retenues par le tamis de la comparaison.

Chaque paire de notices repérée est comparée champ par champ. Chaque élément (paire de champs) a un poids, très fort pour l'ISSN, moins fort pour le nombre de pages.

Et chaque élément a aussi un statut en fonction du cas de figure rencontré (zone absente dans l'une des notices, différence d'une seule lettre dans une chaîne de caractères..). La combinaison de ce statut et de ce poids donne à chaque élément un pourcentage de similarité. Le total divisé par le nombre d'éléments permet d'obtenir le degré de similarité entre deux notices, sa moyenne. Un des principaux obstacles à éviter est de confondre deux volumes différents d'une publication « multivolume », ou une notice décrivant plusieurs volumes avec une autre qui décrit un seul des volumes. *Le numéro de partie* (format *UNIMARC zone 200 $i*) est un renseignement primordial. Cela entraîne Sisyphe à retoucher la conversion des notices provenant de la base **Sibil**, pour distinguer cette sous-zone de celle du titre de partie.

La quadrature du cercle. Tout le problème réside dans le « doigté » du logiciel. Il s'agit de ne pas confondre des notices qui ne décrivent pas le même document. Mais il faut aussi avoir une base utilisable avant quelques dizaines d'années de dédoublonnage manuel. L'enjeu est donc de réduire le nombre de notices à examiner, dites « jumelles dizygotes ».

Trois catégories de notices

1. Les jumelles homozygotes. Certaines notices sont identifiées comme doublons tout de suite : celles qui ont été versées d'une base dans une autre et ont conservé l'ancien numéro d'identification. Il s'agit surtout de notices provenant de **BN-Opale** ou de **Library of Congress** versées dans **Sibil.**

D'autres notices obtiennent un score supérieur à 90 % après la comparaison des différents champs, elle sont aussi semblables, les deux catalogueurs ont rédigé exactement la même notice dans deux bases - ou plus..

Toutes ces notices sont traitées sans intervention humaine, selon une table dont nous parlerons plus bas.

2. Les jumelles dizygotes. Le jeu devient plus difficile : les notices se ressemblent, mais certaines zones diffèrent (une partie du titre, de l'édition, de la collection, l'année de publication varie d'une unité, le nombre de pages est presque identique...) Leur score est entre 70 % et 90 %

Les notices candidates sont marquées par le logiciel, et seule une intervention humaine pourra les sortir de cet état de candidature. L'interface de catalogage présente les deux notices dans deux fenêtres sur le même écran, les zones différentes sont surlignées. Le catalogueur habilité a trois possibilités : il peut décider que les notices sont semblables, auquel cas un traitement *en différé* est déclenché, selon les mêmes paramètres que les jumelles homozygotes ; il peut décider que finalement elles sont dissemblables, et elles perdent leur statut de « candidates ». Enfin, si le catalogueur n'en sait pas plus que le logiciel, on en reste au statu quo, jusqu'à vérification sur les ouvrages eux-mêmes, par exemple.

3. Les non jumelles « pas zygotes ». Les notices qui n'ont aucune concurrente ou bien dont les concurrentes sont trop dissemblables, pour qu'il soit même utile de passer du temps à les examiner, sont chargées dans la base sans autre forme de procès.

Fusionner, pas écraser

Dédoublonner, cela pourrait nous contraindre à annoncer à des bibliothécaires la triste nouvelle que « leur » notice va être « écrasée » par une autre. Une notice conçue avec tant de soin ! Nous avons au contraire opté pour la notice la plus complète possible. Le logiciel de Pica est paramétré de façon à fusionner les notices au lieu de les écraser.

Toutes les données locales ou d'exemplaires sont toujours rattachées à la notice résultant de la fusion dans tous les cas de figure.

Les tables de fusion. Chaque base bibliographique source possède sa table de fusion : il faut choisir, pour chacune des zones de la notice entrante, la conduite à tenir. Le choix est triple : ne rien faire, ajouter, ou remplacer.

Il est bien évident qu'il n'est pas possible de toucher aux données codées sans compromettre la cohérence de la notice. Le titre aussi, sera celui de la notice existant dans la base.

Mais les questions commencent avec la zone de l'adresse. Si la notice résidante n'a pas de zone adresse (tout peut arriver dans une notice de rétroconversion !), il est souhaitable d'ajouter la zone qui a été créée par le catalogueur ayant enrichi cette même notice dans une autre base. Cela ne peut être fait qu'à coup sûr, pour les notices portant un numéro de base source identique (les jumelles homozygotes). De toute façon, une notice sans zone adresse ne saurait être « candidate doublon » sans cette identité de numéro.

En cas de fusion notice A avec notice B : garder notice A et ajouter les zones suivantes de la notice B si elles sont totalement absentes de la notice A : 010*, 011*, 020, 021, 022, 071, 210, 215, 225, 3XX (comparées une par une) [...]

.../...

On voit que nous choisissons d'ajouter les zones de notes. Cela est fait pour ne pas perdre d'information. Mais l'inconvénient est que la même information sera dupliquée si elle est pour une notice dans une note générale (zone 300) et pour l'autre dans une zone de notes dédiée (par exemple 307, note sur la collation). Les entrées aux auteurs sont aussi une richesse supplémentaire. Il serait possible d'ajouter les zones collectivités auteurs aux zones auteurs personnes physiques ou le contraire, à condition bien sûr de ne pas obtenir deux auteurs principaux, mais un auteur principal et un auteur secondaire. Les tests deviennent vraiment trop sophistiqués pour être absolument fiables. Pour éviter cela et garder la cohérence entre la description *ISBD* et les vedettes auteurs, nous prenons le parti de n'ajouter aucune des zones *UNIMARC 7XX*.

Les indexations matière offrent une autre sorte de casse-tête. Il est impossible de tout ajouter sans discernement, et nous ne voulons pas perdre des indexations. Il faut donc prendre une solution intermédiaire, et pour chaque zone, comparer les deux chaînes de caractères. Si pour le même thesaurus le contenu du $a (la « vedette ») n'est pas identique, on ajoute la zone. S'il est identique, on laisse de côté ; car le risque est trop grand d'avoir des indexations redondantes.

Comparer chaque 600 notice A à chaque 600 notice B, chaque 601 notice A à chaque 601 notice B, chaque 602 notice A à chaque 602 notice B, etc.

Si pas de 60X notice A, ajouter systématiquement le 60X de la notice B.

Si 60X $2 notice A >< 60X $2 notice B, ajouter systématiquement le 60X de la notice B.

Si 60X $2 notice A = 60X $2 notice B (ie : 2 rameau, 2 fmesh, etc.) et si 60X $a notice A >< 60X $a notice B, alors ajouter 60X notice B.

Les notices ainsi construites vont être pourvues d'une indexation matière foisonnante : *Rameau, Fmesh, Mesh, Library du Congrès, mots clefs des thèses*... Un affichage bien choisi et des index nombreux rassemblant ou distinguant les différents thesauri permettent à l'utilisateur final et professionnel de ne pas perdre le nord.

Forger des liens

Les accès à ces notices enrichies doivent être - ou devenir - cohérents pour former une base, et non un patchwork de bases. Pour cela, nous avons choisi de relier les points d'accès autres que le titre à des notices autorités[2].

La manière dont ces liens sont effectués par un logiciel ad hoc est décrite dans les spécifications élaborées avec Pica durant tout le 2e trimestre 1999. Chaque type d'autorité - et la zone bibliographique correspondante - est traité par un algorithme approprié. Il faut déterminer quelles sous-zones font partie de la comparaison, et quelles autres ne sont pas concernées. Chaque nouvelle version des spécifications donne lieu à de nouveaux tests. Les liens entre notices de périodiques et entre monographies et collections sont traités dans le même esprit. Un exemple limpide est celui des « titres uniformes ». Prenons la notice autorité titre uniforme « Tracassin et Picarette ». Toutes les zones *UNIMARC 500* (bibliographique) comportant ce titre doivent être examinées, mais certaines sous-zones de la zone 500 doivent être éliminées de la comparaison, comme la sous-zone $k (date de publication) ou $m (langue de la publication). Ces deux sous-zones doivent être exclues de la chaîne à comparer car elles ne font pas partie de la zone correspondante (230) de la notice autorité qui se résume à « Tracassin et Picarette ».

La zone 500 fait partie de la notice bibliographique, la zone 230 appartient à la notice autorité.

500 $a Tracassin et Picarette $k 1999 $m slovène.

230 $a Tracassin et Picarette

Les écueils à éviter sont divers : les collectivités territoriales ou les précisions géographiques dans des langues différentes dans la notice autorité et la notice bibliographique (Turin, Londres, contre Torino et London, voire Londino) peuvent faire échouer les meilleures intentions. La signification du singulier / pluriel dans les zones Rameau fait que l'on ne peut tolérer une lettre de différence dans les chaînes de caractères comparées (« théâtre » ou « théâtres » ne seront pas liés à la même vedette Rameau).

Cerise sur le gâteau, les zones Rameau peuvent comporter un ou plusieurs liens à une ou plusieurs autorités (pour les initiés, le $3 est répétable). Le logiciel de Pica cherche donc d'abord le texte de la zone entière. S'il trouve une vedette Rameau correspondant, le travail est fini. Sinon, la recherche se fait sur les différentes sous-zones :

606##$3027661725$3027229246 $3027346900$2rameau

se traduit par :

606 ##Animaux fabuleux - France - Montpellier (Hérault)

où chacune des sous-zones de la zone 606 est liée à une notice autorité différente. Les seuls animaux fabuleux admis dans le SU étant des carpins[3], la suite de l'histoire vous racontera la chasse aux carpapins, c'est à dire les résultats de toutes ces opérations et le nettoyage de la base.

Mireille Teissèdre
Mireille.Teissedre@abes.fr

1 Cf. *Arabesques* n° 13, Mireille Teissèdre, *Où m'ont-ils mis mon dollar d ?*

2 Cf. *Arabesques* n° 15, Élisabeth Coulouma, *Petit glossaire : qu'est-ce qu'une notice d'autorité ?*

3 Cf. *Arabesques* n° 8, Philippe-Corentin Le Pape, *Boire à plusieurs sources et devenir carpin*

Éplucher, émincer, cuire…

Des notices qui divaguaient en liberté…

Avant le milieu du XXe siècle, les bibliothèques décrivaient les documents selon leurs propres règles. Pendant des siècles, elles ont travaillé avec des inventaires, des listes de titres, plutôt que des catalogues, qui permettent de rassembler les ouvrages d'un même auteur ou sur le même sujet. Depuis quelques décennies, les bibliothèques cataloguent leur fonds selon des règles internationales : en 1971, la publication de la première norme « ISBD (M) » (International Standard Book Description for Monograph Publications) donne aux bibliothécaires un langage commun. Mais si cette normalisation suffisait pour harmoniser des catalogues sur papier, elle ne suffit plus à l'ère de l'échange de données informatisées : il faut aussi que les données aient un format compris de tous les programmes. On ajoute donc à une description qui suit les mêmes règles un format international d'échange. Ce format, l'UNIMARC (pour UNIversal MARC) remonte à 1977, et a été adopté en France comme « format national d'échange » en 1993.[1]

Les bibliothèques peuvent échanger leurs données pour que l'utilisateur ait accès à tous leurs documents ; un grand chantier reste cependant ouvert : la « conversion rétrospective » des notices rédigées sur un support papier (fiches ou registres) en notices dématérialisées au format UNIMARC. Quelques sociétés se partagent le marché, et si les techniques diffèrent – saisie manuelle ou numérisation directe – le résultat obtenu est toujours différent de ce que serait une notice créée directement sur un logiciel de catalogage, selon les règles en vigueur aujourd'hui.

…aux notices génétiquement modifiées

Pour un catalogue informatisé, les notices de conversion rétrospective sont parfois aussi indigestes que la tambouille d'une gargote : elles sont marquées du code « i » (notice partiellement conforme à l'ISBD) ou même « n » (non conforme à l'ISBD) dans le label UNIMARC, en position 18.

En quoi cela est-il un problème de santé pour un catalogue (et pour ses gestionnaires) ? Parce que l'information sur la publication décrite n'est pas bien répartie dans les zones ad hoc, qui n'existaient pas lors de la rédaction de la notice : les notices sur papier ont été saisies telles quelles, ou presque, et « reformatées », c'est-à-dire découpées selon le format UNIMARC.

« Mais encore ? » dira celui pour qui la gestion des catalogues est une sinécure. En fait, les programmes informatiques, bien moins avisés que les bibliothécaires, ne savent pas toujours retrouver les informations : par exemple ils ignorent le nombre de pages d'un document s'il est ailleurs que dans la sous-zone UNIMARC dédiée. Il faut préciser ici que ce format se découpe en zones et sous-zones, où une zone est identifiée par une « étiquette » à trois chiffres, et une sous-zone par un code (une lettre) introduit par le signe dollar « $ ».

Le nombre de pages d'un document devrait se trouver dans la zone **215**, sous-zone **$a**, exclusivement. Et lors d'un chargement, lorsque l'on compare aux notices du Sudoc celles que l'on veut y faire entrer, il faut, pour que la comparaison ait un sens, que les éléments de chaque notice soient dans les zones dédiées du format.

Le défi de l'intégration dans le Sudoc de ces notices de conversion rétrospective non-ISBD conduit à élaborer des recettes originales pour éviter une forme grave d'intoxication : la « doublonnite aiguë ». Le service *Données* de l'ABES utilise à cette fin des moulinettes qui recomposent les notices au moment de la conversion.

Les moulinettes d'importation sont une obligation : le format interne du Sudoc n'est pas l'UNIMARC, mais un format propriétaire, le Pica+. Il faut donc convertir les notices du format national d'échange au format interne. On profite de l'occasion pour rapprocher autant que possible les notices reçues des notices produites actuellement.

Comment savoir où se trouvent la mention de responsabilité, le lieu d'édition, la collation, ces éléments indispensables de l'identification d'une publication ?

Pour y parvenir, il faut mouliner les seuls éléments disponibles : la ponctuation et le vocabulaire.

Éplucher le titre

Retrouver le début du titre et mettre à sa place la mention de responsabilité est la première étape de la recette. C'est en ISBD un élément localisé dans la « zone du titre et de la mention de responsabilité », en UNIMARC la zone **200**.

L'identification du début du titre, qui marque le premier caractère indexé, permet de ne pas avoir de mots parasites qui craquent sous la dent, tel l'insecte dans la salade. Par ailleurs, à l'autre extrémité du titre, les mots qui peuvent introduire la mention de responsabilité, doivent être repérés, pour être les marqueurs de la sous-zone spécifique. Mais la sauce ne peut être liée qu'à l'aide d'une ponctuation décisive et identique d'une notice à l'autre.

Pour rendre ces lignes moins indigestes, on ne décrira pas l'identification des autres éléments de la zone, ni les indicateurs.

Où commence le titre ?

Identifier le début du titre propre a l'air simple et dépourvu de pièges : c'est compter sans le fait qu'en UNIMARC, le titre doit comporter des caractères de non-tri encadrant l'article initial présent. La BNF s'est servi de ces caractères de non-tri d'une façon originale, dans les notices dites « CGI » (Catalogue général des imprimés).

Les fiches saisies commencent parfois par le nom de l'auteur, au lieu du titre. Dans ce cas, l'auteur est encadré de caractères de non-tri, et la moulinette sait où commence vraiment le titre propre.

Par exemple

$a<Nicodème Paillon.**>**Aux origines de la socca[2]

Où **<** et **>** représentent les codes des caractères de non-tri.

Le programme de conversion utilise cette chaîne de caractères « Nicodème Paillon », qui doit être ignorée, pour trouver le premier caractère du titre : il la replace dans la mention de responsabilité, c'est-à-dire dans le **$f**, tout en traitant correctement l'article si besoin est.

Et si la zone ne peut pas fournir de **$f** (si utile quand le titre se résume à « Œuvres complètes »), on peut aller plus loin dans la notice, chercher s'il existe une zone 700 (auteur personne physique) ou 710 (auteur collectivité) ou 720 (auteur nom de famille), et l'y copier, en inversant les éléments du nom de personne.

$a@**Aux origines de la socca**$fNicodème Paillon

…/…

Parfois les notices n'ont aucun caractère de non-tri : il faut alors scruter le premier mot de la zone titre, et si c'est un article à éliminer (présent dans une liste noire) c'est après lui que l'on place le caractère qui marque le vrai début du titre.
Voici la structure « farcie » que l'on rencontre dans un lot de notices de thèses.
$aUniversité de Nice. Faculté de cuisine. La pissaladière, sa préparation, sa cuisson
Ici, il faut tester le premier mot du titre : si c'est « Université », avancez jusqu'au point suivant. Puis regardez le mot qui suit le premier point : si c'est « Faculté », avancez encore jusqu'au point suivant. Cela relève autant du jeu de l'oie que de la moulinette à notices. Et enfin, n'oubliez pas que **La** est sur la liste noire, c'est un article.
Le résultat sera :
$aLa @pissaladière, sa préparation, sa cuisson
où **@** marque le début du premier mot indexé du titre.

Où finit le titre ?

Les mots utilisés par les bibliothécaires pour introduire une mention de responsabilité entrent dans une liste longue mais, heureusement, pas infinie.
En éminçant la notice selon la ponctuation, et en analysant si dans un des fragments obtenus on lit **par** on retrouve la mention de responsabilité.

Deux exemples
La ratatouille dans tous ses états, **par** Charles Nissart
scindée en deux tronçons par la virgule, peut donner

$a**La @ratatouille dans tous ses états**$fpar Charles Nissart
où **$f** marque le début de la mention de responsabilité.
La version ci-dessus est rendue tout à fait digeste pour le catalogue
mais
La ratatouille niçoise par Charles Nissart
sans aucune ponctuation, donnera au mieux

$aLa @ratatouille niçoise par Charles Nissart
où le contenu du **$a** englobe le nom de l'auteur : la ratatouille est trop lourde, la chaîne de caractères fait une longueur double de ce qu'elle devrait faire.

Car imaginez que l'on ait ce titre
La culture de la blette par les paysans de l'an mil, rédigé par Alphonsine Trognon

Si l'on suppose que **par** suffit pour trouver la mention de responsabilité, on obtient
$a**La @culture de la blette**$fpar les paysans de l'an mil, rédigé**$g** par Alphonsine Trognon

Alors que si la ponctuation et les mots clés sont reliés dans l'analyse de la structure de la chaîne de caractères, on retrouve un titre correct
$a**La @culture de la blette par les paysans de l'an mil**$frédigé par Alphonsine Trognon

À cet aromate doux-amer qu'est la ponctuation (absente / présente) s'ajoute une décision à prendre à propos de la casse des mots utilisés comme marqueurs : faut-il les chercher uniquement en minuscules, ou étendre la recherche aux éléments comportant des majuscules ? Seule une analyse poussée du fichier à convertir peut le dire.

Trouver la bonne adresse

Une autre zone nécessaire à l'identification du document est celle de l'adresse. La ponctuation ISBD permet de distinguer « lieu : éditeur, date », que l'UNIMARC codifie **$a**, **$c**, **$d**. Mais dans des notices de conversion rétrospective, on peut trouver un état mixte entre ponctuation et marquage par sous-zones, comme : « lieu, éditeur**$d**date ». Il faut donc pouvoir distinguer les deux premières sous-zones par la présence de la virgule.
C'est trop facile, car un programme informatique balaie la notice dans l'ordre où elle se présente, sauf si on le force à revenir en arrière. Donc, on sait que la première virgule rencontrée introduit l'éditeur. La date sera l'élément qui suit le **$d**. Elle peut être entourée de parenthèses, suivie d'un point, ou d'une date plus précise, tous éléments à identifier et éliminer proprement.
Mais tout se complique, hélas, après la date de publication.
Dans une bonne partie des notices, les éléments de la collation ne sont pas dans la zone prévue (**215**), mais directement à la suite de la date, en **210$d**, comme ceci :
210$a Barcelona, Libreria Bosch**$d**1947. **In-**8°, XII-463 **p.**, fig. et pl.

Prendre une petite collation

Une fois de plus, il faut recourir à l'alliance du vocabulaire et de la ponctuation.

. In- signale le début de la zone de la collation, qui commence par le format, et non pas par le nombre de pages, comme dans les notices standard. Mais ce peut être aussi **. Gr. in-** et pourquoi pas **. Grand in-**. Gare à celui qui voudra faire le malin en réduisant la recherche à **Gr**, chaîne commune aux deux expressions : une sous-zone commençant par **Grenoble** sera traitée comme de la pagination, et non comme le lieu d'édition !
Comment retrouver cette pagination ? Il faut que la rondelle obtenue contienne **p.** Oui, mais si ce sont des colonnes, et pas des pages ? Un test de plus.
Quant aux illustrations, elles sont dans le fragment de zone qui n'entre ni dans la catégorie **pagination**, ni dans la catégorie **format**.
« Il suffit » donc de découper la zone. Mais en fait, il ne faut pas moins de cinq étapes de conversion des tronçons de zones dans des sous-zones intermédiaires pour arriver au bout de la zone de la collation.

. In-8°, XII-463 **p.**, fig. et pl.
devient
215 $a**XII-463 p.**$cfig. et pl.**$d**In-8°

Éplucher, émincer, cuire...
Illustration de l'auteur *M. T.* *teissedre@abes.fr*

Éviter les fausses notes

En continuant le balayage des zones ISBD, on arrive aux notes : UNIMARC zone **300** (note générale), morceau de choix ! On y trouve tout ce qui n'a pas sa place dans le corps de la notice : mention de collection (**225**), titre traduit par le catalogueur (**541**), mention d'extrait d'un périodique (**305**), liste des œuvres contenues (**327**), résumé de la carrière de l'auteur (pourquoi pas **330**), supplément à un périodique (**311**), mention de thèse (**328**), ou encore des notes relatives à un exemplaire, qui n'ont pas leur place dans le corps de la notice d'un catalogue collectif.
Une salade pas très niçoise, cette zone **300** !

Cuire à feu vif

Les méthodes maintenant éprouvées permettent une cuisson à feu vif, sans que les notices attachent au fond du serveur.
Faire gratiner est tout à fait néfaste, car le temps de traitement d'un lot serait augmenté de la durée de récurage du programme.

Servir tiède

Une fois que la recette est élaborée, on fait venir les goûteurs : le mets est présenté sur trois plats, trois formats différents de la même notice (format reçu en entrée, format interne Pica +, format de catalogage du Sudoc). Les trois plats doivent avoir la même saveur, même si l'aspect en est différent. La réaction des goûteurs est essentielle : leur digestion est surveillée une semaine durant. Au moindre signe de malaise, il faut revenir sur le programme de conversion, ajouter un peu de basilic sur la collection, écumer les notes surabondantes, lier la mention de responsabilité.
Ce n'est que lorsque le cuisinier est sûr que les goûteurs ont survécu à l'ingestion des notices que le catalogue est alimenté à son tour.

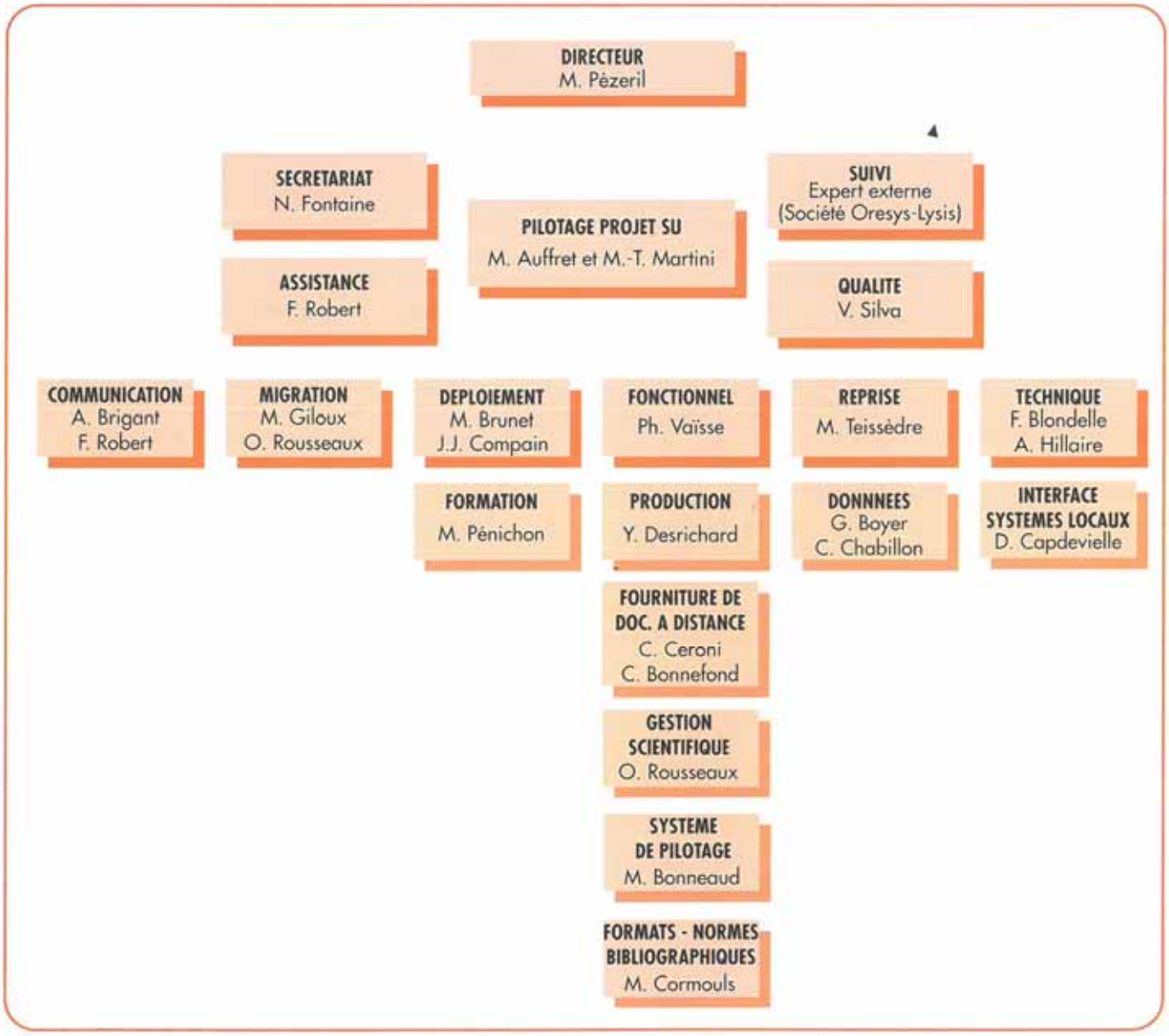
***200 $a**<L'Abbé Desanges.>De l'art de faire des beignets de fleurs de courgettes*
***210 $a**Aspremont, Librairie du Château**$d**1925 (25 octobre). **In-**8°, 25 **p.**, ill. portr.*
***300 $a**Collection Cuisine du terroir*

200$a@**De l'art de faire des beignets de fleurs de courgettes**$fl'Abbé Desanges
210$a Aspremont**$c**Librairie du Château**$d**1925
215$a**25 p.**$cill. portr.**$d**In-8°
225$a Collection Cuisine du terroir

M. Teissèdre
donnees@abes.fr

[1] Pour en savoir plus sur les formats et leur utilisation : www.mbm.org/rencontres_2004/leresche-marc.pdf
[2] Toute ressemblance d'un exemple avec des notices existant ou ayant existé ne serait qu'une amusante coïncidence

Le système universitaire de documentation à l'Abes



Ont contribué à ce n° hors série d'*Arabesques*, Michel Auffret, Marie-Thérèse Martini et Florence Robert (pilotage SU), Martine Brunet et Mireille Pénichon (chantier déploiement), Marianne Giloux (chantier migration), Mireille Teissèdre (chantier reprise), Philippe Vaïsse et Mireille Bonneaud (chantier fonctionnel), Frédérique Blondelle (chantier technique), Annie Brigant (chantier communication) et Sylvette Salvit (Arabesques).

Bonne retraite Mireille

ABES ✆ 04 67 54 84
Arabesques ↔ Sylv
Comité de rédaction
Clientèle ↔ Laurence
Communication ↔ An
Web http ://www.abes.

r de la publication
ggy Pézeril
N 1269-0589
ABES
25, rue Dupuytren . BP 4367
34196 MONTPELLIER CEDEX 5