

L'ABF et le marché des notices bibliographiques

Au congrès de Nantes en mai 1987, un carrefour était déjà consacré au "marché des notices bibliographiques"; trois systèmes y étaient successivement présentés: La base BN - Opale par Mme F. Finelli, LIBRA par M.P. Sanz, la base Electre par M.M. Bony. Un débat riche et d'emblée ressenti comme important s'en était suivi, si bien qu'il fut décidé de le prolonger par une journée d'étude, bientôt tenue le 25 juin 1987 à Paris: Ce sont quatre des interventions entendues ce jour là que nous reproduisons ci-dessous, plus la conclusion de Marcelle Beaudiquez. La motion adoptée le 25 juin sur le sujet a été publiée dans le numéro 136 du bulletin.

La guerre des Marc

par Marc CHAUVEINC, Conservateur en chef du
Centre de prêt de la Bibliothèque nationale

Je me présente ici comme un ancien combattant. Un ancien combattant de la guerre des formats qui a sévi en Europe dans les années 1972 à 1982. Ce fut la guerre d'Intermarc, suivie par la guerre d'Unimarc.

Je croyais ces guerres terminées avec l'arrivée de la télématique et des bases de données en conversationnel. Il semble que non et que le confit larvé resurgisse. Faut-il encore se battre pour un \$a ou un \$c ? J'espère bien que non et que l'on pourra dépasser ce débat pour l'élargir et le dédramatiser. Le PAF (Paysage Automatisé Français) est confus et laisse perplexe la plupart d'entre nous.

Je vais donc essayer de préciser aussi clairement que possible ce qu'est un format, à quoi il sert, et surtout quels problèmes pose son utilisation.

I - QU'EST CE QU'UN FORMAT?

Le format est la structure des données en ordinateur. Par comparaison, c'est une bonnetière dont chaque tiroir contient un seul type de données et porte une étiquette indiquant cette donnée.

Le programme va ouvrir le tiroir 100 et y trouver le nom des auteurs pour toutes les manipulations demandées. S'il ouvre le tiroir 25, il trouvera le titre...

Tout traitement en ordinateur exige un format. Votre numéro de sécurité sociale est un format. Le programme sait que les caractères 2 et 3 constituent votre date de naissance. Il y a donc des formats simples et des formats compliqués, selon les données et les besoins.

Un format peut être en zones fixes, comme le numéro de sécurité sociale et de compte bancaire, il doit être en zones variables pour les données bibliographiques. Mais même de simples zones variables sont insuffisantes pour les données bibliographiques qui nécessitent un codage plus complexe.

En effet : - l'ordre de classement n'est pas toujours conforme à l'ordre alphabétique (La Fontaine, chiffre en caractères romains, suppression de mots vides).

- Il y a des données primaires comme le nom de l'auteur et des données secondaires comme le prénom, le sous-titre ou la fonction.

- On a besoin de codes précisant les caractéristiques du document : livre ou périodique, niveau intellectuel, congrès ou publication officielle.

- On a aussi besoin de préciser à la fois la nature des données : nom de personne ou de collectivité, titre propre, titre parallèle... et leur fonction dans la notice (auteur principal, secondaire...).

II - A QUOI SERT UN FORMAT?

Un format est obligatoire. Sans format, le logiciel ne peut reconnaître qu'un flot informe de caractères parce qu'il n'y a pas de bornes, d'étiquettes ou de drapeaux pour déterminer les éléments significatifs.

On découpe donc les données pour les identifier, les retrouver et les traiter. Ces traitements peuvent être :

- l'indexation, c'est-à-dire créer des index par auteurs, titres, sujets.

- Le tri alphabétique ou numérique des données.

- La sélection de tels types de données (livres, périodiques, thèses...).

- L'impression ou l'affichage de certaines données et pas d'autres.

Dans chacun des cas, on a besoin de reconnaître tel élément comme le nom des auteurs, le titre de la collection, voire le numéro de volume dans la collection. On a besoin, par exemple, de séparer titre et sous-titre si on veut imprimer un catalogue abrégé.

Donc, votre format va dépendre de la nature des données, du découpage et des traitements que vous voulez. Pour la Sécurité Sociale, pour les banques, couper selon le nombre de caractères suffit, cela ne suffit pas pour les bibliothèques.

III - HISTOIRE

Ce fut le mérite de la Bibliothèque du Congrès de créer, en 1968, le 1^{er} format de type MARC (MACHINE READABLE CATALOGING) qui associait différentes techniques pour découper et analyser les données bibliographiques.

Une étiquette à 3 chiffres permet de définir la nature d'une donnée (nom de personne) et sa fonction (auteur principal).

- des indicateurs permettent de signaler un traitement particulier (le 2^{ème} indicateur de format USMARC permet de supprimer les articles initiaux dans le classement des titres).

- Des codes de sous-zones délimitent dans une zone des éléments secondaires (prénom d'auteur et sous-titre).

- Des codes fixes donnent les caractéristiques du document traité (livre, périodique, congrès, livre pour enfant...).

Le même format peut, en conséquence, traiter des documents de types différents. A partir de ce format MARC II, presque tous les pays ont créé leur propre format en modifiant le format original selon leurs besoins : BNB MARC, MONOCLE, INTERMARC, IBERMARC, puis UNIMARC résultant d'un accord international pour l'échange des données entre bibliothèques nationales.

Parallèlement, il y a eu le développement des ISBD (International Standard Book Description). A ce sujet, il faut insister sur un point. L'ISBD n'est pas un format machine. Il est insuffisant pour reconnaître les données et les codes ne sont pas assez significatifs. Il y manque toutes les zones vedettes. Il y manque les codes fixes. La ponctuation ISBD n'est pas suf-

fisamment claire pour un codage machine.

IV - COMPLEXITE DES FORMATS

Il faut bien être conscient que les premiers formats et surtout le premier format INTERMARC étaient destinés à la production de grands catalogues imprimés.

Or, il faut reconnaître d'une part :

- que le classement de nos fichiers ne suit pas l'ordre alphabétique ;

- que pour rendre ces catalogues utilisables, nous avons tendance à faire des regroupements systématiques dans un classement alphabétique (cf. le classement de la Bible ou des oeuvres de Victor Hugo dans un grand catalogue).

On a donc introduit dans le format tous les éléments pour trier correctement les notices selon l'ordre d'un grand catalogue (Saint avant Pape avant roi etc. Oeuvres complètes, choisies, individuelles, titres de classement, transformation des chiffres romains en chiffres arabes etc.). C'était la grande époque des traitements en différé.

Maintenant, nous sommes passés aux catalogues en ligne avec accès direct aux notices individuelles. La recherche par mots du titre, par troncature, les renvois automatiques, toutes ces techniques éliminent bon nombre de problèmes de classement. De plus, il y a peu de chance qu'un gros catalogue puisse être publié désormais. Mais aussi, les règles de classement ont été simplifiées et les problèmes de noms complexes abandonnés plus ou moins.

Le résultat de cette évolution a été une certaine simplification des formats, donc une diminution de l'importance des codes et de leur raffinement avec pour conséquence un rapprochement entre les formats.

Ainsi, INTERMARC a été rapproché d'UNIMARC chaque fois que cela était possible.

V - POURQUOI ENCORE DES DIFFERENCES ?

La réponse n'est pas facile.

- D'abord parce que les formats changent. Ils sont sujets à révisions permanentes, ce qui rend difficile une comparaison exacte.

- Ensuite, parce que les différences sont irrégulières. Par certains points, le format MARC LC est plus complexe, notamment dans la zone fixe, que le format INTERMARC, dans d'autres, c'est l'inverse.

- Ensuite, parce que nos formats sont remplis de différences parfois minuscules, parfois fondamentales, mais même les minuscules sont souvent bloquantes. Désigner un auteur par 700 au lieu de 100, utiliser le \$b à la place du \$m a peu d'importance, la conversion est aisée.

Mettre une virgule à la place du \$m, introduire dans la même zone plusieurs éléments, rajouter plusieurs indicateurs, rend la conversion plus difficile.

Une comparaison sérieuse exige donc une mise en rapport de tous les codes du format en comprenant ce qu'ils veulent dire et en essayant de faire passer les données de l'un dans l'autre. Il m'a fallu plus de 2 ans pour comprendre MARC LC, le traduire et l'adapter dans MONOCLE ; la mise au point d'INTERMARC a dû prendre 2 à 3 ans.

Il s'agit donc d'une étude très détaillée et je ne pense pas que cette étude ait été faite avant le choix de certains formats ni encore moins qu'elle ait besoin d'être refaite.

Mais il y a aussi des problèmes plus fondamentaux. Par exemple, la zone titre d'UNIMARC contient à la fois le titre propre et le titre parallèle alors qu'INTERMARC attribue 2 zones différentes. Dans le 1^{er} cas, l'impression du pavé ISBD est plus simple mais l'indexation nécessite la répétition des titres dans d'autres zones. INTERMARC complique l'édition du pavé ISBD mais simplifie les accès.

Ce sont des choix sur lesquels on peut éternellement discuter.

Il y a aussi une différence fondamentale entre INTERMARC et UNIMARC d'une part, MARC LC d'autre part : Le traitement des niveaux bibliographiques, c'est-à-dire des ouvrages en plusieurs volumes. MARC LC ignore cette notion parce qu'il n'utilise pas les chaînages informatiques. La seule possibilité est de regrouper les niveaux dans une seule notice avec le gros inconvénient de rendre les volumes inaccessibles.

La zone 505, dite zone de contenu, n'est pas indexée. INTERMARC et UNIMARC offrent la possibilité de chaînages informatiques en traitant chaque niveau individuellement, ce qui à la fois offre un accès à chaque titre de volume et à l'ensemble.

Vous trouverez plusieurs exemples dans les documents distribués.

Il est faux de dire qu'INTERMARC a abandonné les niveaux. On a juste trouvé

une technique informatique permettant d'introduire le nom de l'auteur dans les notices de volumes.

IV - CAUSES DES DIFFERENCES

Toutes les différences proviennent de deux causes principales.

Il s'agit d'abord des choix entre le tri et l'impression. Soit on trie comme on imprime, soit on imprime comme on trie, soit on sépare les éléments à imprimer en les répétant ou en ajoutant des codes.

Les services documentaires comme Pascal ou Chemical Abstracts se sont peu souciés de format car, ayant des données simples et homogènes (auteurs modernes), ils pouvaient trier comme ils imprimaient.

La deuxième cause des différences provient des règles de catalogage. Un format est fondé sur des règles de catalogage dont il reflète les choix particuliers.

Le coeur du problème ce ne sont pas les codes mais le contenu des zones. Pour une bonne conversion, le contenu d'une zone doit être identique à celui de la zone correspondante de l'autre format. Or, ce sont les règles de catalogage qui définissent ce contenu.

Même si vous avez la même zone collective en MARC LC et en INTERMARC, dans un cas vous aurez la vedette «US»..., dans l'autre «Etats-Unis»... Comment traduire automatiquement ? Dans la zone auteur, vous avez :

- en USMARC : Lawrence, D.H.

- en INTERMARC : Lawrence \$m David Herbert.

USMARC ne distingue pas le prénom qui n'est séparé du nom que par une virgule alors que le format INTERMARC utilise un code de sous-zone.

Dans ce domaine, les règles de catalogage AACR2 et les règles françaises divergent.

VII - CONVERSION

Beaucoup diront que les différences étant minimes, on peut toujours «convertir». C'est vrai, la conversion de notices d'un format dans un autre est possible, théoriquement. Mais, pratiquement, toutes ou presque toutes les expériences de conversion ont été longues, difficiles, onéreuses. Il y a d'abord des cas nombreux de non correspondance de codes. Une donnée définie dans un format ne l'est pas dans l'autre. Par exemple, la zone 521 de MARC LC, Audience du livre, n'est pas reprise dans INTERMARC.

Il y a surtout des différences de contenu.

Demandez au Centre national du CCN les problèmes non encore résolus de la fusion de 4 fichiers.

Demandez à Montpellier les problèmes rencontrés par l'intégration de MONOCLE dans SIBIL.

Demandez à la BN les problèmes d'intégration de fichiers successifs dans le système GEAC.

Demandez à la Villette les problèmes de conversion de MEDICIS en GEAC.

Sans compter le refus des Américains de traduire INTERMARC en LCMARC malgré l'existence d'UNIMARC. Constatez l'inexistence du Réseau Marc International planifié depuis bientôt 15 ans.

On a créé pour cela le format UNIMARC, tout le monde s'est engagé à l'utiliser pour l'échange de ses données. Qui le fait ? Personne vraiment, car derrière ces problèmes de codes, il y a des problèmes de catalogage, plus difficiles à résoudre dans une conversion automatique.

Cela fait 15 ans au moins qu'on parle d'échange international des données. La France a reçu pendant 10 ans les bandes américaines et les Etats-Unis les bandes françaises : qu'en ont-ils fait ? Rien. Chacun a continué à cataloguer les livres de l'autre.

Les différences paraissent minimes, négligeables tant la structure est identique mais ces différences font que la conversion n'est jamais parfaite et qu'il y a toujours perte d'information.

On doit donc limiter les conversions au maximum.

VIII - RELATIVISER

Il faut aussi relativiser les différences et dire que dans le monde des formats, il n'y a pas de vérité absolue.

Les différences proviennent d'analyses et donc de besoins différents (les Français ont toujours voulu des classements plus raffinés que les Américains). Peut-on en juger de façon absolue ?

De plus, il ne suffit pas d'avoir les mêmes règles et le même format pour être toujours compatible. Le catalogage n'est pas une science exacte.

En effet, le format dépend du catalogage personnel. Il suffit de regarder l'OCLC pour constater que, pour un même ouvrage, il peut y avoir plusieurs notices très différentes, plus ou moins complètes, n'ayant pas forcément les mêmes vedettes.

Enfin et surtout, les formats traduisent seulement la complexité du catalogage.

Le traitement d'un livre simple sera toujours simple et très semblable dans tous les formats. Mais un livre compliqué (Beilstein, comptes rendus de fouilles, Bourbaki) sera toujours compliqué à cataloguer.

Donc, toute la difficulté provient des éditeurs qui créent des structures bibliographiques complexes que nous devons reproduire.

Il faut aussi relativiser parce qu'un format est un ensemble très riche qui prévoit tous les cas de catalogage, y compris les plus difficiles. Ceux-ci, heureusement, ne se rencontrent pas tous les jours. Il y a peut-être certains codes d'INTERMARC qui n'ont jamais été utilisés ! Mais tous les formats prévoient des sous-ensembles et des niveaux de complétude.

Par exemple, dans la base BN-OPALE, une partie des notices provenant de la récupération de fichiers ne sont pas en INTERMARC complet. On a regroupé les pavés ISBD et on a sorti les vedettes et points d'accès qui sont les plus importants.

IX - QUE CHOISIR ?

La situation, comme vous le voyez, n'est pas précisément limpide. On peut cependant dégager quelques certitudes.

- Tout d'abord, il faut un format. Votre ordinateur ne pourra rien vous sortir si tout est en vrac.

- Mais, il ne faut surtout pas créer son propre format, cela ne fait qu'ajouter un étage à la tour de Babel.

- Ensuite, il faut absolument un format de type MARC.

. Seuls ces formats offrent une analyse suffisamment fine des données bibliographiques pour en tirer le maximum. Grâce à ces formats, vous pouvez extraire de votre fichier les livres d'enfants illustrés, les romans policiers ou les congrès scientifiques puis les classer alphabétiquement, par auteurs, ou par titres, par illustrateurs ou numériquement par cote Dewey.

Ils peuvent vous paraître trop complexes au départ, mais sachez qu'une loi informatique affirme que les besoins croissent avec les résultats. Ce qui peut paraître inutile aujourd'hui sera nécessaire demain.

. Ils sont d'une utilisation générale dans les bibliothèques et assurent la compatibilité. Nous allons, c'est certain, vers une circulation courante des données bibliographiques. Tous ceux qui n'auront pas de

format MARC seront exclus de cette circulation.

. Ils sont ensuite extrêmement pratiques et efficaces. Grâce à un découpage précis et bien structuré, vous disposez, dans un minimum d'espace, d'une notice complète et extrêmement simple d'utilisation.

La structure en notice, en zones, en sous-zones, les indicateurs, les codes fixes, offrent toutes les possibilités de tri, de sélection et d'impression.

La rédaction des programmes en est facilitée.

. Dans un avenir très proche, ils vous seront fournis par les grands réservoirs de catalogage que sont les bibliothèques nationales, ou de grands serveurs de type SIBIL, OCLC.

Dans toute nouvelle installation, on doit absolument prévoir, dès maintenant, l'alimentation du catalogage par l'extérieur à plus de 90 %.

On ne pourra jamais se passer de notices bibliographiques mais on devra se passer de les faire.

X - RESTE LA QUESTION : QUEL FORMAT MARC CHOISIR ?

Je maintiens mon point de vue relativiste. Je ne crois plus qu'il faille se battre pour un \$a ou une zone 410. Il y a des différences entre les 3 grands formats utilisables en France : UNIMARC, USMARC et INTERMARC, mais qui peut dire absolument quel est le meilleur ?

Néanmoins, il faut bien choisir UN format pour l'ensemble des systèmes français. On ne pourra développer un réseau national cohérent s'il faut sans arrêt convertir

les données d'un système dans un autre avec chaque fois des pertes d'informations, sans oublier le coût de telles conversions.

Ne pouvant choisir uniquement sur des critères internes, il faut, me semble-t-il, utiliser d'autres critères de sélection qui sont :

- les règles de catalogage ;
- le contexte national ;
- quel est le fournisseur de données le plus utile ?

A la limite, on pourrait dire, choisissez n'importe quel format MARC. Mais il faut se rappeler que le format est à l'image de règles de catalogage et que choisir l'un c'est choisir les autres.

Mais il faut surtout tenir compte du contexte national et de l'existant.

Ce qui a fait la force des Américains et le succès de leurs réseaux, c'est en grande partie le fait qu'ils ont accepté et adopté le format de la Bibliothèque du Congrès. Même le puissant OCLC n'a pas cherché à créer son propre format. Il existait un format. Le monde des bibliothèques américaines l'a adopté sans gaspiller son énergie à discuter de \$n.

En France aussi, INTERMARC existe depuis 1975, créé par un groupe de bibliothécaires français et étrangers, autour de la BN.

Pourquoi avoir, à droite et à gauche, choisi d'autres formats ou même pas de format du tout ? Quelles que soient les raisons invoquées, le résultat est très dommageable pour la coordination prochaine d'un réseau bibliographique national.

Les arguments avancés, qui ne sont pas

techniques, ne tiennent pas devant cette évidence : il y a un format dit national qui vaut ce qu'il vaut, qui peut être amélioré sur certains points et on a choisi autre chose en créant ce PAF (paysage automatisé français), aussi confus que l'autre.

Les arguments en faveur d'INTERMARC sont :

- l'antériorité ;
- les notices bibliographiques françaises sont publiées en INTERMARC ;
- INTERMARC reflète naturellement les règles de catalogage françaises.

Si l'on veut un PAF clair et simple, tous les systèmes locaux doivent pouvoir accepter INTERMARC et si possible l'alimenter.

Reste le problème d'UNIMARC. C'est un format d'échange entre bibliothèques nationales. On peut donc dire que les données étrangères seront reçues en France dans le format UNIMARC et traduites une seule fois en INTERMARC pour être ensuite utilisées par toutes les bibliothèques. Il faut une seule conversion au niveau national de tout ce qui vient de l'étranger pour avoir un réseau français simple et uniforme.

On créera ainsi un seul réservoir national dans lequel chacun puisera selon ses besoins, et qu'il alimentera selon ses possibilités.

Ce réseau bibliographique national doit être notre ambition à tous et nous devons tout faire pour coordonner nos efforts et normaliser nos procédures, afin d'économiser du travail.

Puissions-nous être entendus !