Guerre ou paix ?

Le choix d'un format de catalogage

par Marc Chauveinc

Inspecteur général des bibliothèques

Un format est une « structure caractérisant la présentation des informations au sein d'un ordinateur ». Sur les supports magnétiques, les données sont enregistrées de façon continue, sans délimitations. Pour être lue par le programme de l'ordinateur, toute donnée doit être définie par un format.

En effet, les quatre besoins fondamentaux du traitement de l'information sont : 1. le tri alphabétique ou numérique des données, 2. la sélection d'une donnée, 3. l'indexation, 4. l'impression.

Pour cela, il faut reconnaître les données dans le continuum d'octets. On a donc défini des enregistrements, eux-mêmes divisés en « champs » ou « zones », puis en « sous-zones ».

Les premiers formats étaient en zones fixes. Le numéro de sécurité sociale en est un exemple. Le programme sait que le premier caractère représente le sexe, les deux suivants l'année de naissance, et les traite en conséquence. Les premiers formats bibliographiques ont suivi cette règle où l'auteur se voyait attribuer 10 caractères, le titre 25 et ainsi de suite (*cf.* Agape[1]).

Il y eut ensuite les formats variables avec des « drapeaux » ou codes de zones : A = auteur, B = titre (IPEC). Mais la grande découverte fut le format MARC américain en 1965. En effet, les précédents ne pouvaient rendre compte complètement des données de catalogage (plusieurs auteurs, plusieurs titres, plusieurs volumes, etc.). Ce format a introduit de nouveaux concepts qui ont permis de rendre compte de la complexité des notices bibliographiques :

- pour les tris et les statistiques, on a besoin de données codées en zones fixes ;
- les données bibliographiques sont éminemment variables et aléatoires ;
- le classement peut être différent de la présentation (La Fontaine, chiffres romains, premier article) ;
- la nature du document doit être précisée (livre, périodique, thèse, congrès) ;
- il y a des données primaires (nom), utilisées pour les tris et des données secondaires (prénom, sous-titre), non utilisées ;
- chaque donnée peut être définie par sa nature ou sa fonction dans la notice (nom vedette principale) ;
- les relations entre les éléments de la notice peuvent être complexes (traducteur, titre parallèle, complément à un 2e tome).

Les éléments principaux du format MARC sont :

- le *guide* (24 caractères) qui définit dans des zones fixes des informations générales sur la notice (longueur, date, statut) ;
- des *codes fixes* donnant les caractères principaux du document (livre, date, sujet) ;
- des *zones de longueur variable* et non obligatoires ;

1. Catalogue collectif de périodiques lancé à Nice vers 1971.

– le *répertoire*, sorte de table des matières des zones de longueur variable de la notice ;
– les *étiquettes à trois chiffres* ayant des fonctions précises. Le chiffre des centaines indique la fonction (1 = vedette principale, 6 = vedette matière). Les dizaines et les unités désignent la nature (00 = personne physique, 01 = collectivité, 11 = congrès) ;
– les *indicateurs*, en général à deux chiffres, donnent des informations complémentaires ou servent de codes de tri ;
– les *codes de sous-zones* séparent les éléments à l'intérieur d'une zone (*$a* = nom, *$b* = prénom).

A partir de là, de nombreux pays ont fabriqué leur propre format, en commençant par l'Angleterre (BNB-MARC) et la France (Monocle). Mais, auparavant, les Américains avaient eu la sage précaution de faire définir par l'ISO une norme internationale qui donnera à tous les enfants de MARC la même structure (ISO 2709). Enfin, en 1977, UNIMARC essaya de faire la synthèse en trouvant des compromis entre l'Europe et les États-Unis pour définir un format d'échange. La structure de tous ces formats est identique, mais la codification détaillée de certains formats diffère légèrement.

Pourquoi ? Quatre raisons à ces divergences qui portent, il faut le dire, sur des détails et non sur la philosophie générale du format qui reste la même, sauf pour l'Allemagne qui, ayant des règles de catalogage différentes (RAK), a créé un format différent de la famille des MARC (MAB1).

La première raison se trouve dans les règles de catalogage. Un format traduit ces règles en un langage compréhensible par l'informatique. Cela concerne, notamment, les règles de tri, de présentation des noms, la définition des éléments à prendre en compte, la ponctuation. Malgré les normes internationales, il n'est pas certain que les AACR2[2] soient identiques aux normes françaises. Le catalogage n'est pas une science exacte (*cf.* dans l'OCLC, différentes notices pour le même ouvrage).

La deuxième raison se trouve dans l'objectif que l'on se fixe en définissant un format. Les Américains, en créant l'USMARC, n'avaient qu'un objectif : remplacer leur système de distribution de fiches par la fourniture des notices sur bande magnétique. Donc aucun souci de tri ni de présentation. Par contre, quand les Français ont préparé Monocle puis INTERMARC, leur objectif a été la production de catalogues de différents types, dont le Catalogue général de la Bibliothèque nationale. Il faut rappeler qu'à cette époque, les fichiers en ligne n'existaient pas et que la seule sortie possible était le catalogue imprimé. UNIMARC a été défini comme un format d'échange entre bibliothèques nationales, donc comme un moyen terme entre la richesse des formats européens et la simplicité du format américain, aussi neutre que possible par rapport aux règles de catalogage et à la ponctuation.

La troisième est le choix, ou non, de la flexibilité. Le format peut être neutre et ne contenir aucune règle préétablie ou, au contraire, fournir une codification précise des données pour en conditionner le traitement. Dès le début, les auteurs de MARC ont suivi cette voie que les formats suivants n'ont fait qu'approfondir. Ainsi, dans le premier format MARC, les codes de sous-zones étaient un $ simple. Les Anglais et les Français ont demandé le rajout d'une lettre pour préciser ce séparateur (*$a*) et l'utiliser comme clé de tri.

La quatrième est le choix fait entre la redondance des informations et un traitement plus complexe. C'est-à-dire entre le tri et l'édition. Les Américains ont préféré répéter, si besoin était, l'auteur ou le titre dans leur forme « à imprimer » et leur forme « à classer », alors que les Français ont inséré des codes pour que le traitement soit fait par l'ordinateur sur une même donnée non répétée. Ainsi, UNIMARC met dans la même zone le titre propre et le titre parallèle, alors qu'INTERMARC leur attribue deux zones différentes. Facilité d'impression du pavé ISBD dans un cas, facilité d'accès dans l'autre.

Ces quatre raisons ont conduit les fabriquants d'INTERMARC à multiplier les codes de tri et de sélection (indicateurs pour distinguer les titres significatifs des non significatifs, codes de sous-zones pour préciser les éléments à garder ou à supprimer lors d'édition de listes abrégées, etc.). Il faut rappeler que dans un grand catalogue les problèmes de rangement des notices sont difficiles (auteurs prolifiques, homonymes, œuvres complètes, choisies, extraits, etc.).

Avant de porter un jugement péremptoire sur les formats, il faut rappeler quelques vérités premières :

1. Ce n'est pas le format qui est complexe, c'est l'édition (Beilstein, Bourbaki, ouvrages en plusieurs volumes, collections). Contrairement aux bibliographies d'articles, les catalogues de bibliothèque ne sont pas homogènes (congrès, collectivités, plusieurs volumes, anonymes, etc.). Et les bibliothécaires ont la faiblesse de vouloir fournir le maximum d'accès à un document.

2. Les formats évoluent. Le format USMARC reçoit régulièrement des ajouts ou des modifications. INTERMARC a été modifié juste avant l'installation de BN-OPALE. Même UNIMARC est en cours de révision. Quelle version compare-t-on ?

3. Les différences sont irrégulières et, sans une analyse détaillée, il est difficile de justifier son choix. Certaines sont minimes (désigner l'auteur par 100 ou 700, utiliser *$b* ou *$m* a peu d'importance). D'autres sont plus fondamentales parce que le découpage des données ne suit pas la même logique. Une des plus grandes difficultés se situe dans le traitement des ouvrages en plusieurs volumes qui est soit synthétique dans l'USMARC (une seule notice avec des zones de contenu, inaccessibles dans certains logiciels à l'interrogation), soit analytique dans INTERMARC et UNIMARC (plusieurs notices liées).

4. Ce ne sont pas les codes qui importent mais le contenu des zones. S'il est différent, il rend la conversion parfois difficile ou impossible (*cf.* l'exemple du titre parallèle).

5. Quoi qu'on dise, convertir des données d'un format dans un autre est complexe, onéreux et long (*cf.* le CCN, le Pancatalogue, SIBIL, la Villette). Il vaut mieux l'éviter ou, au mieux, ne le faire qu'une fois.

6. INTERMARC est un format de traitement, et, à ce titre, plus précis. UNIMARC est un format d'échange international, donc plus neutre vis-à-vis des règles nationales. Le bloc des 900 est prévu pour l'usage local et doit être défini par l'utilisateur.

7. Il est impossible de modifier le format d'une grosse base de données, qu'elle soit en USMARC, en BNB-MARC ou en INTERMARC. Quoi qu'on veuille, on est obligé de garder la structure du début.

8. Enfin, trois raisons font que la question du choix d'un format ne se pose plus de façon aussi cruciale qu'autrefois. Les bibliothèques créent de moins en moins de notices et récupèrent celles d'une source bibliographique nationale. Les fichiers sont maintenant tous « en ligne », il n'y aura plus jamais d'édition de grands catalogues avec de graves problèmes de classement, donc les subtilités de la codification sont moins nécessaires. Enfin, les accès par sujets, mots clés ou mots significatifs deviennent

2. *Anglo-american cataloguing rules*, 2d ed.

sans doute plus importants que les accès signalétiques aux vedettes normalisées.

Cependant, cela ne veut pas dire que les précisions doivent disparaître, car l'interrogation booléenne en ligne de gros fichiers (10 à 30 millions de notices) nécessite des index soigneusement codés et délimités. De même, l'affichage de notices abrégées nécessite un découpage précis des zones de données.

Partant du principe qu'il faut un format, et un format MARC, le choix de ce format dans la famille des MARC ne peut plus être absolu et se fait donc maintenant sur des critères externes (fournisseur principal des données, contexte national, règles de catalogage). La grande force des Américains est d'avoir accepté le format USMARC. Tous les réseaux sont donc dans le même format. C'est ce qu'il faut souhaiter à la France. Cette unité est plus importante que les subtilités du *$b*.