Format ? Formats !

ou de l'intérêt de savoir de quoi on parle

par Dominique Lahary
avec le concours de Philippe Raccah

Le format, qui n'était autrefois pour les bibliothécaires qu'un des éléments de la collation et ne troublait que lorsqu'il était *à l'italienne* (dois-je mentionner la hauteur avant la largeur ou l'inverse ?), est un terme magique depuis qu'on s'est avisé d'informatiser. Brandi par les constructeurs, agité par la direction du Livre et de la Lecture et les directions régionales des affaires culturelles, chouchou des cahiers des charges, il est mis à toutes les sauces. Or, l'emploi de ce terme ne va pas sans de fâcheuses confusions, volontaires ou involontaires, de la part des uns ou des autres. Essayons donc de faire le point.

Qu'est-ce qu'un format ?

Qu'est-ce qu'un format en informatique ?

Deux choses. Le *Journal officiel de la République française*[1] a donné une définition officielle du format en informatique : c'est « [l]'agencement structuré d'un support de données » ou « [la] disposition des données elles-mêmes. »

Mais les dictionnaires et lexiques d'informatique donnent parfois une définition plus large, par exemple : « description de la présentation des informations dans la mémoire de l'ordinateur, sur un support de mémoire auxiliaire *ou lors de leur édition sur un périphérique*[2] ».

De fait, ce terme renvoie en informatique à deux notions bien distinctes :
– la structure des données, c'est-à-dire la façon dont elles sont agencées et codées pour pouvoir être identifiées et traitées par un ordinateur ;
– la présentation des données à l'écran ou sur papier.

Qu'est-ce qu'un format bibliographique ?

Quatre choses. Un format bibliographique est un cas particulier de format informatique qui contient des données bibliographiques destinées à être traitées comme telles. On peut en distinguer sommairement quatre sortes :
– le format de stockage : c'est la structure des données dans une base bibliographique, ou plus précisément la structure d'un ou plusieurs fichiers d'une base de données gérée par un logiciel documentaire ou de bibliothèque, qui peut comprendre par ailleurs d'autres fichiers non bibliographiques (emprunteurs, transactions de prêt, etc.) ; cette structure dépend étroitement du

1. Arrêté du 30 décembre 1983 relatif à l'enrichissement du vocabulaire de l'informatique (*JO* du 19 février 1984).

2. *Informatique : dictionnaire* / [sous la dir. de] Yves-Robert de La Villeguérin. – Paris : Éd. La Villeguérin, 1986. – (Les dictionnaires La Villeguérin-Revue fiduciaire).

SGBD utilisé par le logiciel ;
– le format d'échange : c'est la structure des données bibliographiques telles qu'elles sont importées dans un système donné ou exportées vers un autre système ;
– le format de saisie ou de catalogage : c'est la grille de saisie qui apparaît à l'écran lorsque le catalogueur crée une notice, ou la présentation des données telles que le catalogueur les a saisies ; on l'appelle aussi *masque* de saisie, terme informatique qui dit bien ce qu'il veut dire ;
– le format d'affichage et le format d'édition : c'est la présentation de données bibliographiques sur un écran ou sur papier, obtenues à la suite d'une procédure de recherche ou de sélection.

On voit que les deux premières notions sont des notions de structure, alors que les deux secondes sont des notions de présentation. Ainsi, l'ISBD n'est en gestion informatique qu'un format d'affichage, alors qu'en gestion manuelle c'est aussi le « format » de catalogage.

Cette distinction entre quatre sens dérivés des deux significations fondamentales n'est en réalité pas propre aux formats bibliographiques, mais elle est dans ce cas à la fois particulièrement claire et particulièrement utile.

Comment passe-t-on d'un format à un autre ?

De deux façons. Les données bibliographiques peuvent, telles Vichnou, se manifester sous différents avatars ; elles passent d'un type de format à l'autre, voire d'un format à l'autre de même type.

Ainsi, une notice en format d'échange importée par un système est *convertie* dans le format de stockage du-dit système grâce à un *programme de conversion*, vulgairement appelé *moulinette*. A la base d'un tel programme se trouvent des tables de conversion qui constituent un système de correspondance entre la structure de l'un et de l'autre format. Inversement, une notice peut être convertie du format de stockage en format d'échange pour l'exportation. Ces conversions peuvent entraîner des pertes ou des déformations de données (voir encadré).

Mais une notice peut aussi être affichée ou imprimée dans un des formats proposés par le système : il s'agit alors d'une simple réorganisation des données, sans risque de perte d'information. Enfin, lorsqu'on catalogue, les données stockées dépendent directement de celles qui ont été saisies.

Format et normalisation : le monde MARC

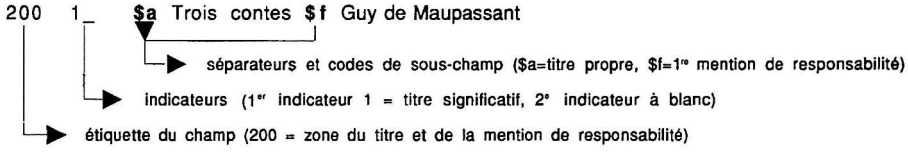
Un format découpe généralement les données en champs identifiés par un code. Ces champs peuvent être en nombre limité et préétabli par le logiciel, ou bien en nombre variable, paramétrable par l'utilisateur, comme c'est maintenant le cas pour de nombreux systèmes de gestion de bases de données. Ils peuvent être de longueur fixe, c'est-à-dire contenir un nombre de caractères limité, défini pour chacun d'eux, ou de longueur variable, au gré des données saisies ou récupérées par l'utilisateur.

Les formats bibliographiques peuvent être structurés en un nombre limité de champs de longueur fixe, mais cette solution n'est guère satisfaisante. L'apport des formats de la famille MARC a été de proposer une structure complexe adaptée aux données bibliographiques.

Ces formats sont normalisés, et non paramétrés au gré des utilisateurs. Ils sont structurés en champs, identifiés chacun par une étiquette composée d'un nombre à trois chiffres et commençant par deux indicateurs numériques permettant d'identifier la nature des informations ou d'opérer certains traitements ou tris.

Un nombre limité de ces champs, destinés à recevoir des données codées, est de longueur fixe. Mais la plupart, contenant des données bibliographiques proprement dites, sont de longueur variable ce qui laisse libre le catalogueur de saisir toutes les informations nécessaires. La plupart des champs sont divisés en sous-champs, commençant par un séparateur et un code de sous-champ indiquant de quelle information il s'agit, ce qui permet un traitement fin des données.

Voici par exemple, en format UNIMARC, le champ correspondant à la zone du titre propre et de la mention de responsabilité de l'ISBD :



Enfin, les formats MARC distinguent clairement les champs descriptifs, qui permettent notamment de reconstituer à l'affichage l'ISBD, et les champs d'accès, le cas échéant normalisés par un système d'index ou d'autorité. Ils constituent des standards, des normes (qu'il ne faut pas confondre avec les normes de catalogage proprement dites). Là où les formats normalisés ont tout leur sens, c'est bien évidement comme format d'échange, et l'utilité d'un format d'échange, de référence n'est pas contestable, quelle que soit par ailleurs la variété des formats de fourniture disponibles sur le marché des notices.

Qu'exiger en matière de format ?

Les utilisateurs ou futurs utilisateurs d'un système exigent couramment de celui-ci qu'il « soit au format X », les fournisseurs de logiciels présentent fréquemment leur produit comme étant « au format X ». Ces expressions n'ont pas de sens car on ne sait pas de quel type de format il s'agit.

Faut-il exiger un format d'échange ?

En matière d'importation, la question semble aller de soi si l'on veut pouvoir récupérer des notices bibliographiques. Mais demander l'importation sous un seul format, par exemple UNIMARC, peut se révéler insuffisant. Des données bibliographiques sont diffusées sous d'autres formats d'échange (US-MARC pour OCLC...). D'autres sont diffusées sous plusieurs formats. Il est alors judicieux de récupérer les données dans le format du réservoir source, plutôt que dans un format d'échange issu d'une conversion. Enfin, quand on a dit UNIMARC, on n'a encore rien dit puisqu'en sus de l'UNIMARC officiel dans lequel aucune notice bibliographique n'est à notre connaissance actuellement disponible en France, la BNF fournit des notices sous deux variantes contradictoires

désignées sous cette appellation (voir article sur les UNIMARC). Il est donc sage d'exiger qu'un système sache importer le ou les formats d'échange correspondant aux réservoirs dont on aura besoin.

Quant au format d'échange dont certaines bibliothèques ont besoin pour exporter des notices, s'il ne doit y en avoir qu'un, on peut actuellement penser que ce doit être « UNIMARC version BN-Opale », sous la norme ISO 2709.

Faut-il exiger un format de stockage ?

La question est rarement débattue. Notamment parce que les constructeurs sont en général peu bavards sur leur format de stockage, à l'exception de ceux qui revendiquent un « UNIMARC natif », ce qui tendrait à faire penser que leur stockage se fait dans une structure identique à UNIMARC (lequel ?) ou en tout cas fort proche.

La question de savoir si un format de stockage doit coller le plus possible ou non à tel format d'échange, qui est susceptible d'évoluer, mérite pour le moins discussion. Elle échappe de toutes façons largement aux utilisateurs, qui disposent de critères externes pertinents pour apprécier le format d'un logiciel : est-ce que les informations importées sont correctement traitées et restituées ? Là est toute la question.

Faut-il exiger un format de catalogage ?

Cette autre question controversée n'appelle pas de réponse univoque. Dès lors qu'on ne confond plus échange, stockage et saisie, la question du format de catalogage peut être une question pratique, s'appréciant en termes de « convivialité », comme disent les informaticiens.

Il existe cependant une nette tendance à la généralisation de la saisie « en UNIMARC », entendez le catalogage sur un masque présentant la structure et les caractéristiques (étiquettes de champs, indicateurs, codes de sous-champs) d'un des UNIMARC, qui tend, en France du moins, à devenir le langage commun des bibliothécaires comme l'était auparavant l'ISBD, au point que les deux manuels de catalogage en UNIMARC qui viennent de paraître en France étaient fort attendus[3].

Mais l'essentiel est de disposer du format de catalogage qui permet la meilleure maîtrise possible de la description et des accès (ce sera selon les cas l'un des formats MARC, avec éventuellement des intitulés en clair, ou un format propriétaire) tout en permettant de créer des notices conformément aux normes en vigueur.

Enfin, le catalogage en MARC demandant une solide formation préalable, on pourra lui préférer pour certains types de bibliothèques une grille plus simple à utiliser.

Faut-il exiger des formats d'affichage et d'impression ?

Cette question est plus simple à aborder. On pourra distinguer :
– un format d'affichage professionnel maximal en MARC et/ou en format de saisie ;
– un format ISBD utilisé soit comme affichage professionnel soit comme affichage public ;
– un format d'affichage public qui peut être l'ISBD ou avec libellés en clair (auteur, titre, etc.), et peut être éventuellement paramétré par l'utilisateur ;
– des formats d'impression en fonction des différentes listes qu'on entend éditer (bibliographies, listes d'acquisitions, etc.).

Il n'est pas interdit de penser que si les bibliothécaires sont particulièrement attachés au format ISBD, ce n'est certainement pas le cas du public. Mais, en matière de format d'affichage public, chaque système ou chaque utilisateur de système paramétrable invente actuellement sa solution, ce qui entraîne une fâcheuse dispersion.

Conclusion

La distinction entre les différents types de format est indispensable à la fois pour pouvoir apprécier un logiciel et pour pouvoir utiliser l'information bibliographique disponible, l'essentiel étant que le système local restitue convenablement les informations importées et permette à la fois une description normalisée et des accès et traitements pertinents.

Les risques de la conversion

Convertir une notice bibliographique peut entraîner des pertes ou des déformations d'informations. On peut sommairement distinguer trois types de risques, présentés ici dans le cas de figure d'une conversion par un système importateur dans son format de stockage (que nous appelerons format cible), de notices importées dans un format d'échange.

• Le format cible ne comporte que des champs de longueur fixe.

Si le système dispose d'un tel format non conforme aux formats MARC, les données importées sont alors susceptibles d'être tronquées.

Exemple : une bibliothèque d'une petite commune disposant d'un système sommaire avec champs de longueur fixe pourra importer des notices de la BDP qui la dessert, mais tous les caractères dépassant la longueur de ces limites disparaîtront.

• Le format cible ne prévoit pas un type d'information contenu dans le format d'origine.

Le système importateur peut alors soit ignorer ces informations (elles sont alors perdues) soit les stocker dans un champ (ou un sous-champ) « fourre-tout » prévu à cet effet. (elles sont alors inexploitables), soit enfin les traiter d'une façon qui les rende exploitables, en « rusant » avec le format cible.

Exemple : lors de la conversion en USMARC de notices en UNIMARC, les champs de liens ascendants ou descendants d'UNIMARC (461 et 463) posent problème car ils n'ont pas d'équivalent en USMARC.

• Le format cible prévoit des éléments d'information qui ne sont pas distingués par le format d'origine dans le format d'échange.

Il est parfois possible grâge à un programme de conversion complexe d'identifier les élémets d'information exigés par le format cible.

Exemple : lors de la conversion d'USMARC en UNIMARC, le champ du titre et de la mention de responsabilité, qui peut ne comporter que trois sous-champs en USMARC, est décomposé par programme dans les sous-champs exigés par UNIMARC en utilisant la ponctuation de l'ISBD.

3. *Cataloguer en UNIMARC : un jeu d'enfant* / Philippe-Corentin Le Pape. – Équinoxe : Fédération française de coopération entre bibliothèques, 1993 ; et *UNIMARC : Manuel de catalogage* / Marie-Renée Cazabon. – Paris : Éd. du Cercle de la librairie, 1993. – (Collection Bibliothèques). Ces deux excellents manuels ne portent que sur le catalogage des monographies et des publications en série.

Mais, le plus souvent, il y a déformation, un champ ou un sous-champ du format cible devenant le réceptacle d'informations plus larges que celles qui correspondent à sa définition exacte.

Exemple : on peut sur le CD-ROM ELECTRE récupérer des notices en format ELECTRE, en UNIMARC, en LC-MARC ou en UK-MARC. Mais c'est le format ELECTRE qui est le format de la base du Cercle de la librairie, les notices étant converties par le fournisseur dans trois formats MARC. Or il s'agit d'un format « maison » nettement moins riche que les formats MARC. Par exemple, il ne distingue pas entre auteur principal et auteurs secondaires. Il en résulte que la conversion de notices ELECTRE dans un format MARC, fut-elle proposée par de Cercle de la librairie lui-même sur son propre CD-ROM, ne peut pas offrir plus d'informations qu'il n'en existe dans le format d'origine.