Enjeux et défis
d'un système de catalogage multilingue et multiécriture

L'un des objectifs prioritaires de la BULAC, dès sa création en 2002, a été de se doter d'un SIGB. L'opération a été prévue en deux temps. Pendant la phase de préparation de la nouvelle bibliothèque, l'accent serait mis sur la constitution d'un catalogue informatisé commun aux composantes qui se regrouperont dans le futur bâtiment. Puis, dès que possible, s'y adjoindraient un module de gestion des acquisitions et un module de gestion des périodiques. La seconde phase, plus ou moins concomitante avec l'ouverture du nouvel établissement (2007-2008), serait consacrée à la mise en place d'un module de gestion des lecteurs et de communication des documents et au développement d'un système d'information complet. Réunir en les déménageant des collections de tailles diverses et couvrant les différents champs de l'orientalisme avec plus ou moins de redondances exige de les identifier de la manière la plus efficace possible, c'est-à-dire à l'aide d'un outil informatique. Or, sur ce plan, la situation actuelle de chacune des bibliothèques composantes est extrêmement contrastée : aucune n'a de SIGB propre, à l'exception du fonds slave de la Sorbonne ; quelques-unes signalent dans le Sudoc au moins une partie de leurs collections (la plupart du temps, celles en caractères latins, comme à la BIULO, mais parfois aussi certaines en caractères non latins, par le biais de notices romanisées, comme pour le fonds slave de la Sorbonne ou les fonds d'Asie du sud-est de la BIULO depuis 1989 environ ou tous les fonds de cette même bibliothèque depuis 2001 seulement) ; enfin, de nombreux fonds ne sont toujours pas signalés autrement que dans des catalogues sur fiches traditionnels (les fonds anciens ou la majorité des fonds en écritures non latines de la BIULO, fonds du centre d'études slaves, etc.). Quelques initiatives avaient cependant déjà été prises pour cataloguer informatiquement certaines collections en caractères non latins : fonds chinois, japonais, coréens (CJK) et en arabe et persan de la BIULO (dans OCLC, depuis 2000, en parallèle avec le catalogage dans le Sudoc – en caractères latins seulement), fonds CJK de la maison de l'Asie (dans le logiciel AGATE, localement). Bref, un vaste chantier de mise en ordre des données catalographiques existant sous forme informatique devait être rapidement entrepris, grâce à ce SIGB commun, sur lequel pourrait aussi s'appuyer le non moins vaste et redoutable, et tout aussi urgent, chantier de rétroconversion de toutes les notices «papier». Ce projet de catalogue possède une caractéristique tout à fait spécifique : celle de réunir des données relatives à des documents non seulement en de nombreuses langues, mais aussi en de très nombreuses écritures, et d'avoir pour objectif de les présenter simultanément à ses utilisateurs, le cas échéant, à la fois en romanisation et en graphie originale. La réalisation de cet objectif n'est pas sans poser de nombreux problèmes, que je vais examiner maintenant sommairement, sous trois aspects : bibliothéconomique, technique et culturel.

1. Enjeu bibliothéconomique

L'enjeu bibliothéconomique du nouveau catalogue multilingue[1] et multiécriture de la BULAC est double : d'une part, **assurer la fourniture des informations bibliographiques dans les langues et écritures d'origine** des documents mis à disposition du public ; de l'autre, **fournir simultanément et *systématiquement* en caractères latins toutes les informations bibliographiques *quelle que soit l'écriture d'origine du document décrit*.** La raison du premier objectif (accès dans les langues et écritures d'origine) d'ailleurs conforme aux normes de catalogage en vigueur[3], est évidente en ce qui concerne la langue, et théoriquement aussi en ce qui concerne l'écriture : seule la description bibliographique fidèle d'un document permet de le retrouver sans problème majeur. En outre, il est naturel de permettre aux spécialistes de telle ou telle langue orientale d'accéder à l'information bibliographique qu'ils recherchent sous sa forme native, tout comme il est nécessaire de fournir aux locuteurs de ces langues un accès à cette même information sans qu'ils soient obligés d'utiliser un système de romanisation ou de recourir systématiquement pour cela à l'aide de bibliothécaires. Il est donc prévu que le système soit capable d'offrir des données catalographiques dans autant d'écritures non latines (mais aussi latines étendues) que de besoin, à plus ou moins long terme selon les langues. On commencera par le CJK, l'arabe et le persan (déjà catalogués à la BIULO dans OCLC en romanisation et en caractères originaux) ; on devrait y ajouter le thaï ainsi que les notices des ouvrages en écritures cyrillique, grecque, arménienne et géorgienne, dès qu'elles pourront être gérées dans le Sudoc (voir ci-dessous). Le second objectif (accès en romanisation[4]), pour être moins bien compris, n'en est pas moins impératif, voire prioritaire. Sa justification majeure est purement pragmatique : la BULAC est d'abord une bibliothèque appartenant à l'environnement culturel occidental, où l'écriture latine est naturellement omniprésente, et elle s'adressera fondamentalement à un public utilisant quotidiennement cette écriture[5]. En outre, la diffusion de l'écriture latine sur la planète étant ce qu'elle est, au moins dans certaines sphères de la vie soumises à d'intenses échanges internationaux, un catalogue présentant des données en caractères latins permet maintenant de toucher, grâce à Internet, d'autres publics, situés n'importe où dans le monde. Les autres raisons sont plus théoriques, bien qu'avec de nombreuses implications pratiques. D'abord il serait dommage de revenir au temps des fichiers manuels qui imposaient de faire une série de recherches successives par écriture pour retrouver des références en grec, français, anglais, russe, arabe, persan, lorsque l'on s'intéressait à Aristote par exemple : un système moderne doit permettre d'exécuter en une seule opération une recherche, notamment par auteur ou par sujet[6], couvrant des références en différentes écritures. Ce type d'accès est d'ailleurs tout à fait conforme à la philosophie sous-jacente à la constitution des fonds tels que seront ceux de la BULAC et que sont déjà ceux de la BIULO : ouverture et multidisciplinarité. Il s'agit bien, dans un tel type de bibliothèque, de rendre accessible un maximum de documentation sur tous les sujets liés à l'orientalisme, quelle que soit la langue dans laquelle elle est écrite : ouvrages en russe sur le Vietnam, en japonais sur l'Iran, en allemand sur l'Afrique, en arabe sur l'Espagne, etc. : tout ce qui concerne un sujet doit être directement repérable indépendamment de la langue ou de l'écriture, sans *a priori* ; c'est ensuite au lecteur, et à lui seul, de faire son choix, en fonction de ses compétences, et non aux spécialistes de telle ou telle langue au nom d'une connaissance dont ils seraient les seuls détenteurs : un ouvrage en chinois doit pouvoir être repéré et cité en *pinyin* seul par un lecteur

Interprétation du premier type

«C'est ce qui justifiera de **ne pas différencier *c* et *ç* en turc** ou **en azéri latin** dans les tris, mais de **distinguer en persan** dans le même cas, ب (b)(U+0628) de پ (p)(U+067E), **parce qu'en arabe déjà,** ب (b) ne se distingue par exemple de ت (t) (U+062A) que par les points. D'autres cas sont plus problématiques […]. Aux spécialistes de trancher…»

V. H. hachard@idf.ext.jussieu.fr

non sinisant, mais néanmoins capable de consulter des illustrations ou une bibliographie qui lui seront profitables. Le bibliothécaire ne peut en aucun cas lui dénier l'accès à ce type d'ouvrages : il doit au contraire favoriser ces recherches croisées, même si elles restent marginales. Ensuite, l'offre d'accès aux écritures originales sera forcément limitée par les coûts liés aux matériels informatiques plus ou moins spécifiques à mettre en œuvre : coût directement financier bien sûr, mais aussi coût en espace (il n'est pas imaginable d'équiper tous les catalogues ou tous les postes de travail de la batterie de claviers qui seraient nécessaires à la consultation ou à la saisie de données en écritures latine, grecque, cyrillique(s), géorgienne, arméniennes, arabe(s), indiennes, chinoise, etc.). Et cela sans parler des accès distants que j'évoquais plus haut ; que sait-on de l'équipement informatique de la personne qui consultera dans quelques mois ou dans quelques années le catalogue de la BULAC : un Mac ou un PC ? un récent ou un ancien ? équipé des polices nécessaires ? et ainsi de suite… Enfin, dernier argument, et non des moindres, l'interopérabilité du futur SIGB avec des systèmes extérieurs passe par les caractères latins : on ne peut en effet préjuger, du moins dans la phase actuelle, que chacun des partenaires source ou destinataire des données catalographiques de la BULAC soit à même de manipuler des caractères non latins : une transposition en caractères latins (y compris en Unicode, voir plus loin) des écritures originales garantit dans tous les cas la possibilité de travailler sur des données codées dans des jeux de 256 caractères maximum et en n'ayant besoin que de polices de caractères latines (riches ou même limitées). En ce qui concerne la BULAC, cela est particulièrement vrai, puisqu'à terme, selon le schéma habituel, l'ensemble du catalogage se fera dans un Sudoc devenu Unicode, avant qu'il n'y ait redescente des données vers le système local de la BULAC. Les notices produites ou modifiées par des catalogueurs de la BULAC seront nécessairement en caractères originaux et en romanisation, puisque les bibliothèques qui seraient amenées à se localiser sur ces notices n'ont pas ou n'auront pas (au moins dans un avenir proche) de système local à même de prendre en compte les caractères non latins, voire le codage Unicode[7] (voir plus bas).

2. Enjeu technique

L'usage des caractères latins étendus et non latins (CLENOL) qu'implique l'enjeu bibliothéconomique est rendu possible de façon relativement simple par le codage de caractères Unicode, utilisé de façon native par le SIGB retenu pour la BULAC[8]. Développé par le consortium Unicode, Inc.[9], créé en 1991, ce système de codage informatique des caractères s'est définitivement synchronisé à partir de sa version 1.1 (1993) avec la norme ISO 10646-1 développée elle par le groupe ISO/IEC JTC1/SC2/WG2[10] fondé en 1984. Il est destiné à remplacer les différents systèmes de codage existants, qu'ils soient sur 7 ou 8 bits[11] (ISO 646 = ASCII, ISO 8859 et ses 10 variantes ; dans le monde des bibliothèques, ISO 5426, ANSEL) ou sur 2 x 7 ou 8 bits (pour les caractères CJK : JIS X 0213, GB 2312, Big5, KSC 5601) et en est aujourd'hui à sa version 4 (2003) [= ISO 10646:2003 *Universal Multiple-Octet Coded Character Set (UCS)*]. La plupart des écritures employées aujourd'hui dans le monde, ainsi que quelques écritures mortes ou inventées, y sont codées, occupant 96 382 positions (numérotées en hexadécimal de 0x0 à 0x10FFFF) sur les 1 114 112 possibles[12] Il serait trop long de décrire ici les principes de fonctionnement d'Unicode. On en retiendra simplement un, qui est qu'à chaque entité **abstraite** d'écriture (le «*caractère*», angl. «*character*») désigné théoriquement par un nom (en anglais ; les noms des caractères CJK n'en sont en fait pas puisqu'ils sont du type CJK UNIFIED IDOEGRAPH CHARACTER-*n*, où *n* est le numéro du caractère !), est associé un code numérique (la «*valeur scalaire*», angl. «*code point*»). La valeur scalaire et le nom sont des éléments normatifs ; en outre, un ensemble de propriétés est associé à chaque caractère (casse, sens de l'écriture, chasse, etc.). Un caractère peut recouvrir différentes variantes concrètes nommées *glyphes*[13] ; en réalité, le plus souvent pour des raisons de compatibilité, il existe un certain nombre de variantes (donc théoriquement des glyphes) qui sont encodées, et si les diacritiques flottants ont leur code, de nombreuses combinaisons caractère de base + diacritique(s) ont aussi le leur propre (on parle alors de caractères précomposés). Unicode prescrit ou recommande également un certain nombre d'éléments techniques très importants : tris, sens de l'écriture, etc. Pour un SIGB tel que celui de la BULAC, Unicode en tant que tel (système de *codage des données* dans la base de catalogage) représente donc une condition nécessaire mais non suffisante pour obtenir un système multilingue multiécriture fonctionnel. En effet, ces données n'ont aucune utilité si on ne peut pas les manipuler : en ajouter, les modifier, les supprimer. Pour ce faire, un certain nombre d'outils et de processus informatiques capables de traiter des données Unicode doivent être mis en œuvre. D'abord, au niveau de l'interface homme/machine, ces données doivent être **rendues visibles** : c'est le rôle des polices de caractères adéquates, pour l'affichage à l'écran et l'impression, que ce soit localement ou à distance ; cela suppose des polices largement répandues, ou, au minimum, lorsqu'il s'agit d'accès distant, téléchargeables et installables par l'utilisateur sans problème technique (utilisation indépendante des plates-formes : Windows, Mac, Linux) ou juridico-financier (polices libres de droit ou gratuites) ; de plus, pour certaines écritures, elles doivent posséder un minimum d'«intelligence» (gestion des ligatures ou des variantes contextuelles, positionnement correct des diacritiques, etc.). Il faut aussi, dans le cas où sont mélangées des écritures de sens différents, que le SIGB soit capable d'en tenir compte efficacement en ce qui concerne la mise en page : l'entrelacement de texte de gauche à droite et de droite à gauche est en effet un des défis qu'un système multilingue doit relever. Elles doivent être aussi, bien évidemment, **manipulables** : c'est là le rôle qu'ont à remplir les claviers, matériels ou virtuels. Selon les besoins, il s'agira donc – s'il en est décidé ainsi – de proposer sur des postes publics des claviers matériels différents (biécriture : latin/cyrillique, par ex., ou bien monoécriture, arabe, par ex.) ; des claviers virtuels[14], notamment pour la saisie des diacritiques utilisé dans les romanisations, ou des logiciels de saisie du CJK (comme l'*Input method editor*, IME, fourni par Microsoft), plus facilement généralisables, devront aussi équiper les postes afin de permettre l'utilisation du clavier standard français pour la saisie de tous les CLENOL ; il en existe déjà un certain nombre livrés avec MS Internet Explorer 5+ ou Windows 2000+, mais ils souffraient jusqu'à récemment de la

.../...

limitation gênante de ne pas être configurables[15]. En tout état de cause, les «Tables de caractères», que ce soit celles proposées par Windows ou par un fournisseur de SIGB ne sont pas utilisables autrement qu'occasionnellement (donc certainement pas pour le catalogage). Ces deux aspects de l'accès aux données dépendent aussi d'un autre critère : celui de la distinction poste de travail professionnel et poste de consultation publique (local ou distant), ce qui recouvre plus ou moins la distinction entre phase de production du catalogue et de consultation du catalogue. Par ailleurs, de façon interne, les données Unicode doivent être à même de subir les deux opérations essentielles dans une base de données en général, et dans un SIGB en particulier : l'indexation et le tri. Des outils sont fournis par le standard Unicode et par les documents techniques afférents (*Unicode standard annexes, Unicode technical standards, Unicode technical reports*), l'un des plus importants dans ce domaine étant l'UTS 10, *Unicode collation algorithm,* qui définit un algorithme de positionnement (*collation*) des caractères Unicode les uns par rapport aux autres, permettant d'obtenir un tri par défaut de chaînes de caractères Unicode quelconques (DUCET, *Default Unicode collation element table*). En aucun cas le pur ordre numérique des codes de caractères ne peut servir pour ce classement («*…the only way to get the linguistically-correct order is to use a language-sensitive collation, not a binary collation*», UTR 10, v. 4.0, § 1.8) ; il faut au contraire au minimum adopter l'algorithme défini par Unicode, ou mieux, l'adapter à ses propres besoins, ce qui pose un autre type de problème dans une base multilingue *et* multiécriture (voir ci-dessous). Enfin, il ne faut pas perdre de vue qu'il y a intérêt à permettre une recherche à l'aide d'une chaîne de caractères «appauvrie» (c'est-à-dire dépouillée de tous ses diacritiques), comme c'est déjà habituellement le cas, même s'il est peut-être souhaitable, dans certains cas, de permettre la recherche au moyen de la chaîne de caractères exacte (ou «riche»), tandis que les **réponses à ces recherches doivent toujours être typographiquement riches,** c'est-à-dire avec les caractères pourvus de tous leurs diacritiques et avec les ligatures orthographiques (obligatoires). La seule différence dans les réponses sera, éventuellement, au niveau de leur classement : précis si la requête a été faite sur le mode «riche», global si la requête a été lancé en mode «pauvre». On le voit, toutes ces caractéristiques demandent de la part des fournisseurs de SIGB bien d'autres investissements qu'une simple implémentation d'Unicode pour le stockage des données, même si trop souvent ils ont tendance à se défausser sur le système d'exploitation ou sur des solutions tierces qu'ils laissent le soin à la bibliothèque de trouver, notamment pour les polices ou les claviers.

3. Enjeu linguistique et culturel

Si l'on admet que les problèmes précédents ont été résolus, il reste un dernier aspect, non négligeable, dans la mise en place d'un SIGB multilingue et multiécriture ; celui, précisément lié à sa nature, d'être au croisement de multiples cultures, et d'être donc soumis à la tension qui naît, d'une part, de la nécessité d'être, autant que faire se peut, « neutre», et, d'autre part, des attentes et des exigences de ses utilisateurs, aussi bien «nationaux» que spécialistes français et «occidentaux» de telle ou telle langue. J'illustrerai ce problème par deux exemples : celui de l'ordre des écritures dans les index ; celui des ordres «alphabétiques» nationaux, en liaison notamment avec le problème des «diacritiques». L'ordre des écritures à l'intérieur d'un même index est à première vue simple : il suffirait de prendre l'ordre numérique des caractères Unicode pour régler la question. Or, on l'a vu à l'instant, cet ordre ne convient pas, selon les spécifications même d'Unicode (UT 10) ; en outre, il suffit de regarder les tables pour s'apercevoir que l'on trouve par exemple des caractères grecs dans la section U+0374-03FB, puis, après des sections consacrées au cyrillique, à l'arabe, à l'hébreu, aux écritures de l'Inde, au thaï, à l'éthiopien, etc., dans la section U+1F00-1FFE : un minimum de réorganisation est donc nécessaire pour les tris. Comme la table de positionnement des éléments est adaptable, et qu'il est même recommandé de l'adapter, il faut déterminer un ordre des écritures. Le plus logique est de faire passer l'écriture latine en premier, à la fois pour les raison évoquées ci-dessus (§ 1), et aussi parce que la majorité des données seront nécessairement dans cette écriture, puisque, rappelons-le, il est prévu de doubler systématiquement les champs en écriture non latine d'une romanisation. Pour l'ordre des autres écritures, la question reste ouverte. On pourrait choisir de garder l'ordre de première apparition dans Unicode ; d'adopter un ordre lié à l'importance relative des fonds dans la base (à quel moment ?) ; enfin, - c'est la solution pour laquelle je penche personnellement, on pourrait classer les écritures selon leur type (voir plus bas). L'important, dans tous les cas, est de retenir que cet ordre, s'appliquant à une base multi-écritures, ne peut être que **conventionnel,** et n'implique aucune hiérarchie particulière entre les différentes écritures. Avec les ordres «alphabétiques» nationaux, qui posent aussi la question du traitement des «diacritiques» dans les tris, nous nous trouvons à nouveau confrontés aux contradictions générées par la nature multilingue du SIGB de la BULAC. Les exemples sont bien connus : *ö* est classé traditionnellement en français (quand il apparaît) au même endroit que *o*, tandis qu'il équivaut à *oe* en allemand et constitue en suédois une lettre située **après** z ! Autant dans un catalogue purement national, il est compréhensible de respecter ce genre de critère de classement, autant cela est inconcevable dans un catalogue dont les données sont internationales et dont les publics le sont aussi. La solution est donc de n'utiliser pour le classement général[16] que les caractères de base (caractères «appauvris»), sans diacritiques, avec peut-être, dans les cas où il y a plusieurs diacritiques, un sous-classement en fonction de leur nombre[17]. Au passage, on devra distinguer les véritables diacritiques, ceux qui ont été inventés au fil du développement et de l'expansion d'une écriture (comme la cédille, l'accent, le tilde, le *háček*, l'*ogonek*, etc.) de ceux qui sont nés en même temps (ou pratiquement en même temps) qu'une écriture et en sont constitutifs (comme le point dans l'écriture arabe) : c'est ce qui justifiera de ne pas différencier *c* et *ç* en turc ou en azéri latin[18] **dans les tris,** mais de distinguer en persan dans le même cas, ب (b)(U+0628) de پ (p)(U+067E), parce qu'en arabe déjà, ب (b) ne se distingue par exemple de ت (t) (U+062A) que par les points. D'autres cas sont plus problématiques, la différence entre devanagari क (ka) (U+0915) et क़ (qa) (U+0958) (je penche dans ce cas pour une interprétation du premier type) ou celle entre amharique ተ (ta) (U+1270) et ቸ (cha) (U+1278) (ici, je suspends mon jugement !).

Толстой, Лев Николаевич (1828-1910) -- Biographie -- Famille

Tolstoï, Lev Nikolaevitch (1828-1910) -- Biographie – Famille

Tolstoj, Lev Nikolaevič...

Aux spécialistes de trancher… On notera enfin que certaines pratiques spécifiques d'ordonnancement alphabétique ne peuvent de toutes façons pas être retenues : ainsi le classement par racines, habituel dans la lexicographie arabe. Face à ce type de problèmes ou d'interrogations, ce qui compte, c'est de pouvoir justifier avant tout sur le plan bibliothéconomique des choix qui pourront heurter à l'occasion tel ou tel utilisateur dans ses habitudes culturelles ou dans ses pratiques de recherche. Qu'ils soient bibliothéconomiques, techniques ou culturels, les enjeux et défis du SIGB dont sont en train de se doter la BULAC et les bibliothèques qui la composent, sont, on l'a vu, nombreux et parfois contradictoires. Mais la véritable première que constituera la réalisation du projet sera, à n'en point douter, à la hauteur des efforts qu'il aura fallu fournir.

Types d'écritures, romanisation, translittération, transcription

Les données présentes dans le SIGB de la BULAC couvrent de nombreuses écritures. En gros, elles se répartissent en quatre cas.

1) ***Écriture alphabétique*** *:* écriture utilisant un caractère, simple ou diacrité, ou un groupe de caractères pour un son, que ce soit une voyelle ou une consonne. Exemples : écriture grecque, écritures cyrilliques, écriture géorgienne…

2) ***Écriture consonantique*** *:* écriture utilisant un caractère ou un groupe de caractères pour noter une partie seulement des sons d'une langue, normalement les consonnes et parfois une partie des voyelles. Exemples : écritures hébraïques, écritures arabes (arabe, persan, ourdou…).

3) ***Écriture syllabique*** *:* écriture utilisant un caractère (ou parfois plusieurs fondus en une seule ligature) pour noter une syllabe (le plus souvent du type consonne + voyelle, mais aussi voyelle seule, ou voyelle + consonne, etc.). Exemple : écritures éthiopiennes (amharique, tigrigna, guèze…), écritures indiennes (devanagari, goudjarati, tamoul, etc.), syllabaires CJK (hangul, kana), syllabaires aborigènes du Canada…

NB : c'est ce type d'écriture qui est le moins souvent pris en compte dans les SIGB multiécriture.

4) ***Écriture idéo(phono)graphique*** *:* écriture utilisant un caractère (simple, diacrité ou composé) pour noter une notion[19] exprimée par un ensemble phonétique plus ou moins complexe. Les caractères n'entretiennent aucune relation particulière avec les éléments phonétiques constitutifs (voyelle, consonne, syllabe) de l'expression des notions. L'écriture idéographique est parfois utilisée conjointement avec l'écriture syllabique (comme en japonais par ex.). Exemples : chinois, japonais, coréen.

Rappelons les définitions en vigueur au sein du comité technique ISO/TC48 «Information et documentation», sous-comité SC2 «Conversion des langues écrites» et valides également dans les normes AFNOR s'appliquant au même domaine. On les trouve dans la première partie de toutes les normes élaborées par ces instances, par ex. dans la norme NF ISO 9 (juin 1995), *Translittération des caractères cyrilliques en caractères latins.*

*«La **translittération** est l'opération qui consiste à représenter les caractères d'une écriture alphabétique ou syllabique par les caractères d'un alphabet de conversion. En principe, cette conversion doit se faire caractère par caractère ; chaque caractère du système graphique converti est rendu par un caractère et un seul de l'alphabet de conversion, ce qui est la façon la plus simple d'assurer la réversibilité complète et sans ambiguïté de l'alphabet de conversion dans le système converti.»*
(NF ISO 9:1995F, § 2.2.).

*«La **transcription** est l'opération visant à noter la prononciation d'une langue donnée au moyen d'un système de signes d'une langue de conversion*
[…] La conversion n'est pas strictement réversible.
La transcription peut être utilisée pour la conversion de tous les systèmes d'écriture. Elle est la seule méthode utilisable pour les systèmes non entièrement alphabétiques ou syllabiques et pour toutes les écritures idéophonographiques, comme le chinois.»
(NF ISO 9:1995F, § 2.4.).

*«Pour la **romanisation** (conversion d'écritures non latines dans l'alphabet latin), on peut utiliser soit la translittération, soit la transcription, soit un mélange des deux méthodes, suivant la nature du système converti.»*
(NF ISO 9:1995F, § 2.5.).

On notera que l'usage de digraphes («digrammes») dans les translittérations n'est pas interdit par les normes ISO, bien que soit affirmé le principe de la conversion caractère par caractère (*cf.* ci-dessus). L'avantage de l'utilisation d'un caractère de translittération diacrité est son univocité, son inconvénient l'éventuelle difficulté qu'il y a à saisir des caractères parfois très inhabituels (comme dans la partie de NF ISO 9:1995F concernant les langues non slaves ou le *t* surmonté d'un tréma dans NF ISO 233-2:1993F, *Translittération des caractères arabes en caractères latins*).

À l'inverse, les digraphes sont généralement d'un usage typographique plus simple, mais présentent l'inconvénient de nécessiter dans certains cas des complications de mise en œuvre (ainsi la norme ALA-LC de 1997 est contrainte de distinguer le digraphe *sh* de la combinaison *s* + *h* au moyen d'un signe prime : *s'h*)[20]. En combinant ces définitions avec les types d'écriture définis auparavant, on obtient les possibilités suivantes de réversibilité :

Type d'écriture	réversibilité	
	romanisation → écriture originale	écriture originale → romanisation
Écriture alphabétique	oui (translittération)	- oui (translittération)
Écriture consonantique	oui	- oui, si translittération stricte - non, si transcription (dite «translittération simplifiée» dans les normes AFNOR)
Écriture syllabique	- oui - non, si combinée avec une écriture idéographique (par ex., japonais)	- oui, si translittération stricte - non, si transcription, même partielle (par ex. éthiopien, si l'on note la gémination) - non, si combinée avec une écriture idéographique (par ex., japonais)
Écriture idéographique	non (transcription)	non (transcription)

.../...

C'est en fonction de la réversibilité d'une écriture par rapport à sa romanisation que peuvent être développés des outils de conversion *automatisés* tels que celui qui existe déjà dans le logiciel *OCLC Arabic* actuellement utilisé à la BIULO. C'est donc un point très important, car le doublement des zones en écritures originales par leur romanisation constituera un gros travail pour les catalogueurs. En tant que bibliothèque française, la BULAC se doit de donner la priorité aux normes de romanisation AFNOR et, à défaut, aux normes ISO : c'est le cas pour le cyrillique, le grec, l'arménien, le géorgien, l'arabe, l'hébreu… Si une norme AFNOR ou ISO n'est pas disponible, elle recourt de préférence aux grandes normes internationalement reconnues dans le monde des bibliothèques, notamment aux normes de la Bibliothèque du Congrès (ALA-LC), qui ont en outre l'avantage d'être facilement accessibles et qui reprennent dans certains cas des normes extrêmement répandues : ainsi pour le chinois (*pinyin*), birman, khmer, laotien… Dans certains cas, différents critères ont fait préférer une norme qui aurait dû normalement ne pas être retenue : serbe (translittéré à l'aide de l'orthographe officielle du croate, selon l'usage courant depuis l'époque yougoslave, et non ISO), japonais (ALA-LC = Hepburn modifié, au lieu d'ISO, inutilisée en pratique), devanagari (ALA-LC et non ISO, trop récente et n'en différant que sur quelques points), persan (système propre conforme aux principes ISO, plutôt qu'ISO ne respectant pas paradoxalement ces principes). Enfin, des compromis plus importants ont parfois été nécessaires ou sont en cours d'examen suite à des discussions entre bibliothécaires et spécialistes des langues : ourdou, éthiopien, etc. Dans tous les cas de figure, il faut se souvenir que les systèmes de romanisation adoptés en bibliothèques **ne sont destinés ni à un usage didactique ni à un usage linguistique.** Ils n'ont d'autre finalité que de favoriser la communication et l'échange des données bibliothéconomiques, de la façon la plus automatisable possible et de la façon la plus neutre possible (c'est-à-dire sans interpréter les données d'origine et donc en respectant le principe ISO de l'indépendance de la romanisation par rapport à la langue[21]. Néanmoins, il est évidemment des cas où d'autres contraintes interviennent (importance d'un catalogue existant dans un système donné, lisibilité de la romanisation, etc.), qui obligent au compromis.

V. Hachard
hachard@idf.ext.jussieu.fr

1 Le multilinguisme dont il s'agit ici n'est pas celui de l'interface (qu'elle soit professionnelle ou surtout publique, locale ou distante), mais bien celui des données catalographiques.
2 *Cf.* les recommandations du groupe de travail sur le *Catalogage des documents en caractères non latins,* rédigées par D. Duclos-Faure, MEN, 2002 : http://www.sup.adc.education.fr/bib/Acti/fcnl/titre.htm.
3 «Dans les zones 1 [titre et mention de responsabilité], 2 [édition], 4 [adresse bibliographique] et 6 [collection] l'information est donnée dans la langue ou l'écriture de la publication [….] L'information relative aux zones 5 [collation], 7 [note] et 8 [ISBN et prix] est donnée dans la langue et l'écriture de l'établissement de catalogage, sauf le titre original et les citations dans les notes.» Z 44-050:1989, *Catalogage des monographies,* § 0.6.
4 Sur la romanisation, voir texte ci-dessous.
5 Cette constatation n'est en rien un jugement de valeur, puisqu'*a contrario* il serait parfaitement légitime d'imaginer (et cela est parfois réalisé) des SIGB ou des catalogues qui offrent systématiquement des données bibliographiques originellement en caractères latins (mais pas seulement) transposées en caractères cyrilliques, arabes ou chinois, par exemple, à l'intention de publics dont c'est l'écriture quotidienne.
6 L'indexation matière s'adressant d'abord au lectorat du lieu où se trouve la bibliothèque, elle doit naturellement se faire en français. Un corollaire est que les nom propres étrangers utilisés dans une indexation matière *ne doivent pas* figurer (ou ne doivent pas figurer *seulement*) en caractères originaux (si ceux-ci ne sont pas latins), mais en romanisation afin qu'un lecteur puisse effectuer une recherche sur n'importe quel sujet : il est inconcevable de proposer une vedette du genre *Толстой, Лев Николаевич (1828-1910) -- Biographie -- Famille* au lieu de *Tolstoï, Lev Nikolaevitch (1828-1910) -- Biographie – Famille* (ou, si l'on préfère, *Tolstoj, Lev Nikolaevič* etc.) : comment le lecteur ignorant tout du russe peut-il retrouver un ouvrage en français sur ce sujet ? En revanche, à l'ère de la mondialisation, il est tentant d'imaginer, pourquoi pas ? un système qui propose parallèlement une indexation matière en anglais. On m'objectera que les auteurs comme les sujets sont contrôlés par des autorités dans lesquelles on peut parfaitement stocker les formes originales et les formes romanisées (et c'est d'ailleurs ce qu'il est prévu de faire) ; cela est vrai : mais on admettra cependant que si celles-ci renvoient sur celles-là, il ne sera guère commode pour le lecteur «ignorant» de parcourir les index qui l'intéressent (et encore, j'ai pris un exemple en cyrillique !).
7 C'est d'ailleurs pourquoi l'ABES prévoit de continuer de proposer une exportation des données en ANSEL ou en ISO 646 (ASCII) + ISO 5426:1983 (*Extension of the Latin alphabet coded character set for bibliographic information interchange*)
8 La capacité à utiliser Unicode était l'une des conditions essentielles formulées dans le cahier des charges.
9 Il s'agit donc d'une entité privée, réunissant pour l'essentiel des industriels de l'informatique, mais aussi des vendeurs de SIGB comme VTLS, Sirsi, Ex-Libris et Innovative Interfaces, ainsi qu'OCLC.
10 International standardization organisation / International electrotechnical commission Joint technical committee / Subcommittee 2 / Working group 2.
11 Fournissant respectivement $2^7 = 128$ ou $2^8 = 256$ possibilités d'encodage.
12 Ces positions sont organisées en 17 plans constitués chacun de 256 lignes de 256 cellules, soit 17 x 2^8 x 2^8 (= 17 x 256 x 256). L'essentiel des codes se trouve dans le plan 0, *Basic multilingual plane* (BMP), mais aussi dans le plan 2, *Supplementary ideographic plane* (SIP) (caractères employés en cantonais notamment).
13 Il s'agit là d'à peu près la même différence qu'entre le phonème et le son en linguistique.
14 Il s'agit de pilotes logiciels qui réassignent aux touches d'un clavier matériel des valeurs différentes de celles qui y sont représentées.
15 Microsoft a en effet mis à disposition de tous un outil de modification des pilotes de clavier sur sa page http://www.microsoft.com/globaldev/tools/msklc.mspx. Mais qui dit s'il ne suivra pas un jour le chemin de la police MS Arial Unicode, qui était citée partout comme *la* police Unicode et était disponible encore naguère gratuitement, mais n'est plus livrée maintenant qu'à l'achat des produits Microsoft ?
16 On ne peut en effet s'interdire *a priori* de constituer des sous-ensembles du catalogue dont les index seraient soumis à des règles nationales particulières.
17 Et de même, le classement des lettres modifiées après la forme d'origine : par ex. cyrillique Ӯ (U+04AF) après У (U+0443).
18 Le même raisonnement sera appliqué aux caractères modifiés autrement que par diacritique, comme le ə de l'azéri, regroupé avec le *e*.
19 En gros, un «mot», d'où une autre appellation de ce type d'écriture : «logographique». Cela dit, le système est en réalité plus compliqué.
20 Il faut signaler à ce propos que les normes ALA-LC, contrairement à ce que l'on pense parfois, utilisent *à la fois* des digraphes *et* des caractères diacrités, et même la combinaison des deux (!), y compris parfois pour la romanisation d'une seule et même langue (ainsi dans la translittération de l'arabe *sh* pour ش [ISO/AFNOR : *š*] , mais ش pour ﺷ [comme ISO/AFNOR].
21 «Dans le cas *où un même caractère est utilisé dans deux langues différentes écrites dans le même alphabet, ce caractère doit être translitteré de la même façon, sans tenir compte de la langue à laquelle il appartient»* § 3.2 de la partie préliminaire des normes ISO (par. ex., NF ISO 233-2:1993, *Translittération des caractères arabes en caractères latins).*

La police Times New Roman convient pour les quelques exemples en arabe, en dévanagari et une partie de ceux en cyrillique ; pour l'autre partie de ceux-ci (n. 15) et un exemple en azéri (n. 16), Lucida Sans Unicode ; enfin, pour l'éthiopien, a été utilisée la police Ethiopia Jiret. *V. Hachard*

Vincent Hachard dirige le service général de la bibliothèque interuniversitaire des langues orientales.
BIULO ✆ 01 44 77 87 21 87 30 4 rue de Lille 75007 PARIS