Constitution d'un annuaire de chercheurs internationaux et catalogage dans un SIGB : les problèmes liés à l'intégration d'informations multilingues

Hélène Rousselot

CONSERVATOIRE NATIONAL DES ARTS ET METIERS

INSTITUT NATIONAL DES TECHNIQUES DE LA DOCUMENTATION

MÉMOIRE PRÉSENTÉ EN VUE D'OBTENIR

LE DESS EN SCIENCES DE L'INFORMATION ET DE LA DOCUMENTATION SPÉCIALISÉES

par Hélène ROUSSELOT

Constitution d'un annuaire de chercheurs internationaux et catalogage dans un SIGB : les problèmes liés à l'intégration d'informations multilingues

Mémoire soutenu devant un jury, composé de

Hélène KUTTLEIN

Jean-Max NOYER

Date : octobre 2003
CYCLE SUPERIEUR PROMOTION XXXIII

Sommaire

Remerciements

Patrick Couët (ministère de la Défense)

Claire Ebersold (Ever Team)

Eïtan Grossfeld (BnF)

Vincent Hachard (BIULO)

Françoise Hours-Richard (BIULO)

Nourredine Lamriri (Ever Team)

Chantal Miel (BIULO)

Nathalie Rodriguez (BIULO)

Maud Thenin (Ex Libris France)

Sara Yontan (BnF)

Introduction

Ce mémoire s'appuie sur le travail effectué dans le cadre de mon stage au Centre documentaire du Campus de Défense (ministère de la Défense). Ma mission a consisté, dans un premier temps, à élaborer un questionnaire en vue de l'acquisition d'un système intégré de gestion de bibliothèque (SIGB). L'analyse des réponses faites par des éditeurs de logiciels et des suggestions de questions complémentaires à leur soumettre sont fournies dans la troisième partie de ce mémoire. Dans un second temps, ma mission a été consacrée à la recherche de ressources documentaires dans le domaine de la défense, en langue russe, essentiellement.

Initialement, l'objectif de cette recherche était de recenser les publications périodiques des centres de réflexion stratégique en CEI. Puis elle a été étendue à celles des spécialistes des questions de sécurité et de défense, en CEI et en Europe centrale, afin d'alimenter une base de données, intitulée "Annuaire des chercheurs de défense". Elle figurera parmi les produits proposés aux usagers du Centre documentaire.

Cette mission m'a donc conduite à m'intéresser aux questions du traitement informatique des caractères cyrilliques sur la Toile, dans un SIGB, ainsi qu'aux questions du traitement de l'information en russe qui devra être restituée aux utilisateurs du Centre documentaire du Campus de Défense. En effet, ceux-ci seront en droit de trouver, le plus simplement et le plus rapidement possible, par exemple, le nom d'un chercheur russe ou l'intitulé d'un périodique, spécialisé dans un domaine qui les intéresse. Or le nom de ce chercheur ou de ce périodique en russe peut être recherché à l'aide de graphies différentes. L'enjeu final est par conséquent l'augmentation des probabilités de bonnes réponses aux questions des usagers du Centre documentaire.

Le cas de la langue russe et de son écriture ne couvre pas toutes les questions posées par l'emploi des caractères non latins dans un système informatique. Il n'est pas le plus complexe puisque cette écriture possède des caractéristiques communes avec l'anglais qui a été la première employée dans les systèmes informatiques. En effet, à l'instar de cette langue, le russe s'écrit de gauche à droite, il est alphabétique et distingue les majuscules des minuscules. Ces caractéristiques sont importantes pour la question du codage des caractères.

Mais cette apparente et relative simplicité cache néanmoins des difficultés. Aussi les investigations menées et les conclusions tirées constituent un éclairage particulier de l'ensemble des questions que soulève l'introduction de caractères non latins dans un système informatisé. Du reste, parmi les documents utilisés pour l'élaboration de ce mémoire, nombreux sont ceux qui exposent une démonstration et une réflexion menées à partir d'une langue donnée pour aboutir à des conclusions plus générales.

Les différentes étapes de la chaîne documentaire qui ont été ainsi étudiées, sous l'angle des problématiques posées par les caractères cyrilliques, sont :

- la gestion des caractères cyrilliques et diacrités dans un environnement Windows (logiciels bureautiques Word, Excel et un logiciel de base de données Access),
- la romanisation des caractères cyrilliques,
- le catalogage,
- les systèmes intégrés de gestion de bibliothèque.

Il est assez rapidement apparu que les problèmes ou les interrogations se réduisent, en grande partie, à l'enjeu que représente l'unification des codages informatiques des caractères. Cet enjeu repose dans un premier temps sur l'implantation généralisée d'Unicode dans l'ensemble des systèmes informatiques, qui a déjà débuté. Et comme le montrent les réponses d'éditeurs de systèmes intégrés de gestion de bibliothèques à une grille d'analyse, un consensus universel pour l'adoption d'un jeu de caractères unifié est en passe de s'imposer.
Parmi les autres problèmes qui subsistent, relevons le choix de la romanisation utilisée notamment dans les normes de catalogage.

Les problèmes soulevés par la présence d'écritures cyrilliques dans un centre de documentation ne sont pas aussi anecdotiques qu'il pourrait paraître au premier abord et ils concernent nombre de bibliothèques puisque, d'après l'enquête menée par le ministère de l'Éducation nationale auprès de 104 bibliothèques universitaires, « les ouvrages en cyrillique constituent en nombre les fonds d'ouvrages en caractères non latins les plus importants présents dans les bibliothèques universitaires : 41 %. Ce sont aussi les plus répandus : un tiers des établissements interrogés possède des ouvrages en cyrillique » (11).

L'introduction de caractères non latins dans un système informatique intéresse le Centre documentaire du Campus de Défense qui souhaite développer son fonds en langues étrangères et proposer des produits documentaires élaborés à partir de sources, elles aussi, en langues étrangères.

I. Le Centre Documentaire du Campus de Défense

Le projet du Campus de Défense s'inscrit dans le cadre du réaménagement du site de l'École militaire, et de sa réorientation vers l'enseignement et la recherche, envisagé depuis l'année 2000, par le ministre de la Défense d'alors, Alain Richard.

Ce projet consiste à effectuer un rapprochement des différents pôles d'information du ministère de la Défense (centres de documentation et bibliothèques des centres de formation) pour constituer, dans un premier temps, un centre documentaire de référence, puis un pôle de recherche et enfin un centre de conférences.

Une des finalités de ce grand projet consiste à réformer largement l'attitude des étudiants qui suivent un enseignement militaire, vis-à-vis de la recherche documentaire. Aussi les grands axes retenus par les promoteurs du projet Campus sont :

- la mise en place d'un lieu ouvert de documentation et d'information,
- la création d'un lieu d'observation des progrès de la politique européenne de défense, de ses ramifications institutionnelles, politiques et industrielles,
- la création d'un réseau entre différents organismes de recherche publics et privés en matière de défense,
- l'amélioration de la diffusion du savoir-faire de la recherche dans l'enseignement militaire supérieur et rénovation de l'enseignement de défense,
- l'accès à des activités de recherche pour le personnel du ministère de la Défense,
- la mise à disposition d'outils documentaires modernes.

Le centre documentaire sera ainsi placé dans un réseau de bibliothèques de référence du ministère de la Défense, comme celles du Service historique des Armées, de l'École polytechnique et de bibliothèques extérieures à l'instar de celle de la Fondation Nationale des Sciences Politiques (FNSP) et la Bibliothèque nationale de France (BnF). Ce qui permettra des connexions de bases de données entre le futur Centre documentaire, la FNSP, la BnF et des universités parisienne et lyonnaise.

Le fonds du Centre, spécialisé dans le domaine de la défense et les questions de stratégie et de géopolitique est d'ores et déjà estimé à 200 000 ouvrages. Le Centre est appelé ainsi à s'insérer dans les réseaux des pôles associés sous la direction de la BnF, des CADIST[1], au réseau informatique SUDOC[2], et enfin au réseau américain MERLN[3].

Parmi les futurs services proposés par le Campus, sont envisagés :

- la constitution de dossiers documentaires et dossiers de presse qui seront accessibles via Internet ou Intranet,
- un service de veille en liaison avec le pôle recherche du Campus,
- un signalement des acquisitions sous forme de catalogue,
- des créations de bases de données (de rapports et de traités, notamment),
- un service expert de type « SVP »,
- un « service de recherche poussée d'informations pour des utilisateurs sélectionnés » sur le modèle de celui proposé par la NDU[4],
- une formation à la recherche documentaire,
- une mutualisation du service « Revue de presse » avec la FNSP qui est à l'étude.

Ce nouveau lieu sera ouvert, c'est-à-dire qu'il limitera au maximum les contraintes d'accès aux documents et à l'information, notamment en supprimant la notion de comptoir derrière lequel se trouvent le documentaliste et le bibliothécaire, et en utilisant un mobilier modulaire permettant un réaménagement facile. La

[1] CADIST : Centre d'Acquisition et de Diffusion de l'Information Scientifique et Technique
[2] SUDOC : Système universitaire de documentation
[3] MERLN : Military Education Research Library Network.
[4] NDU : National Defence University

configuration, envisagée en libre accès, pose de nombreuses questions matérielles d'organisation de la bibliothèque, comme celle des codes à barres collés sur les ouvrages. La structure des nouveaux codes à barres sera revue et l'apposition d'un antivol à ces codes sera étudiée.

La mise en place d'un libre accès, voire d'un libre service dans le futur centre documentaire du Campus de défense, sur le site de l'École militaire est un des objectifs du projet, avec la création d'un service de type SVP, mentionné plus haut.

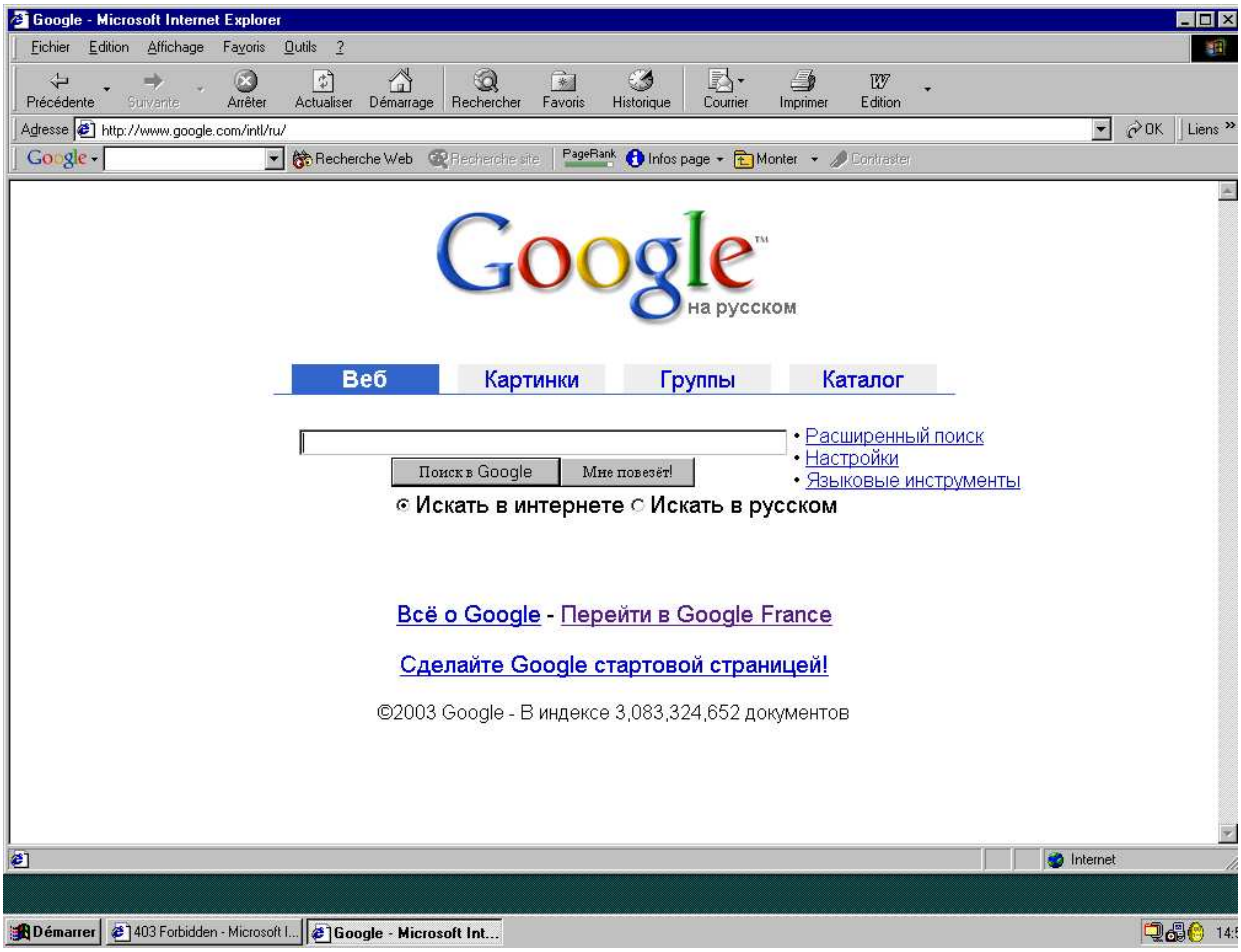
II. Les limites de mon travail de réalisation du mémoire

Ce travail a été effectué dans le cadre d'études des techniques documentaires, auxquelles venaient s'ajouter des compétences culturelles et linguistiques acquises par ailleurs dans le domaine du russe. Cet ensemble de compétences est différent de celui d'une informaticienne et a orienté la façon de traiter ce sujet. Ainsi, le détail des questions de programmation ne sera pas abordé dans ce mémoire.

Par ailleurs, les outils informatiques et Internet, notamment, ont constitué des limites ; l'absence de datation de documents trouvés sur la Toile en est un exemple. La recherche de documents destinés à alimenter ma réflexion sur ce sujet m'a conduite en effet à remarquer que de nombreux textes ne portent pas la date de leur rédaction et / ou de leur mise en ligne. Cela est très gênant dans le cas d'une étude comme celle-ci, étant donné l'évolution très rapide des technologies et des normes.
Aussi, il arrive que certaines assertions présentées dans des textes paraissent dépassées. Un exemple est fourni par le document intitulé « L'édition des textes multilingues » d'Abdel-Malek Boualem et Stéphane Harié, consulté le 3 juillet 2003 sur le site http://rist.cerist.dz/ArticlesFullText/8-1/htm/BOUALEM.htm. Les références du texte les plus récentes n'étant pas postérieures à 1996, on peut en déduire que le texte date de 1996, 1997 au plus tard. Mais il n'est fait mention nulle part de la date de sa parution sur la Toile.

Une autre limite de l'outil Internet réside dans l'impossibilité de retrouver certaines pages en interrogeant de nouveau un moteur de recherche avec la même adresse. La disparition de pages de la Toile peut être un élément limitatif de ce travail de recherche.

FIGURE N°1: INTERFACE DE GOOGLE EN RUSSE



Première partie : La recherche de sites Web et l'alimentation de « l'Annuaire des chercheurs de défense »

La méthode suivie pour recueillir l'information repose essentiellement sur la recherche de sites et sur la navigation de site en site, facilitée par les liens hypertextes. Cette première partie comporte la description de cette méthode et des techniques qui rendent possible cette façon de recueillir l'information. Une fois l'information collectée, elle a dû être traitée et les enjeux de ce traitement sont exposés à la suite.

I. La navigation de site en site

Le point de départ de la recherche des premiers sites russes a été l'étude réalisée, en 1999, par ma responsable de stage sur les centres de réflexion de défense. Puis, la recherche d'information nécessaire à l'accomplissement de ma mission s'est poursuivie grâce aux sources de la Toile et dans une moindre mesure, à partir de sources écrites comme le Courrier des pays de l'Est – la documentation Française.

J'ai, ainsi, recherché les sites de revues, recensées par le Courrier des Pays de l'Est, dans la rubrique « la revue des revues ». Le dépouillement de quelques-uns de ses numéros a suffi à obtenir une liste d'une vingtaine de revues en anglais et en français. Néanmoins, la recherche des sites de ces revues n'a pas été très fructueuse puisqu'un certain nombre d'entre elles n'en possèdent pas (« Europe-Asia Studies », « Politique étrangère », Est-Ovest, East-European Politics and Societies, Contemporary Review...), tandis que d'autres sites ne sont accessibles que sur abonnement (Communist and Post-Communist Studies, Eastern European Economics). Le Courrier des Pays de l'Est m'a également fourni le titre de quelques organes de presse et parfois leur adresse sur la Toile ainsi que des noms de chercheurs dont les articles ou leur traduction en français ont été publiés dans cette revue.

Toutes ces informations ont été des points d'entrée sur la Toile et qui, au moyen de liens hypertextes m'ont permis de découvrir d'autres sites.

Au cours de cette navigation sur la Toile, il est apparu que certains instituts et centres sont difficiles à localiser. Lorsqu'ils sont cités par d'autres, ni leurs coordonnées, ni leur URL ne sont mentionnés. La recherche via un moteur, à l'aide de leur seule dénomination, n'aboutit pas toujours à des réponses. Dans le cas où aucune réponse n'a été trouvée, ils ont été saisis dans la base « Annuaire des chercheurs de défense » tels quels, sans coordonnées. Un exemple est fourni par les départements rattachés à des instituts de l'Académie des sciences de Russie. Certains de ces instituts sont introuvables sur le site de l'Académie, doté d'un moteur de recherche, interrogeable en russe, uniquement.

Les journalistes indépendants, sélectionnés comme experts sur une région, sont, eux aussi, difficiles à localiser : ils sont cités dans des pages Web, sans adresse, ni numéro de téléphone ; ne figure au mieux que le nom des organes de presse

auxquels ils collaborent….. Ce cas des journalistes indépendants pose aussi la question de la présentation de leur production : comment recenser leurs articles ? Choix a été fait de ne recenser que le nom des organes de presse dans lesquels ont été publiés leurs articles.

La recherche des sources via Internet s'est effectuée essentiellement grâce à Google, moteur de recherche créé en 1998 par Larry Page et Sergey Brin (né à Moscou), deux étudiants en doctorat de l'université californienne de Stanford.

II. La recherche via un moteur : l'exemple de Google

Le choix du moteur de recherche Google se justifie par sa prétention à l'exhaustivité et ses performances. Et s'il n'est pas exhaustif, il est le plus complet. Par ailleurs, sachant que la majorité des moteurs interrogent Google au cours de leurs recherches, il semble aussi pertinent d'utiliser directement ce dernier.

L'aspect « minimaliste » de cet interface et l'accès simple à ses outils linguistiques ont aussi joué en sa faveur. En effet, l'interface Alta Vista est moins ergonomique dans sa présentation des outils linguistiques qui demande de cliquer sur « paramètres » puis sur le lien « Langues de recherche » (25 langues sont proposées).

II.1. Les outils linguistiques d'une interface de moteur de recherche

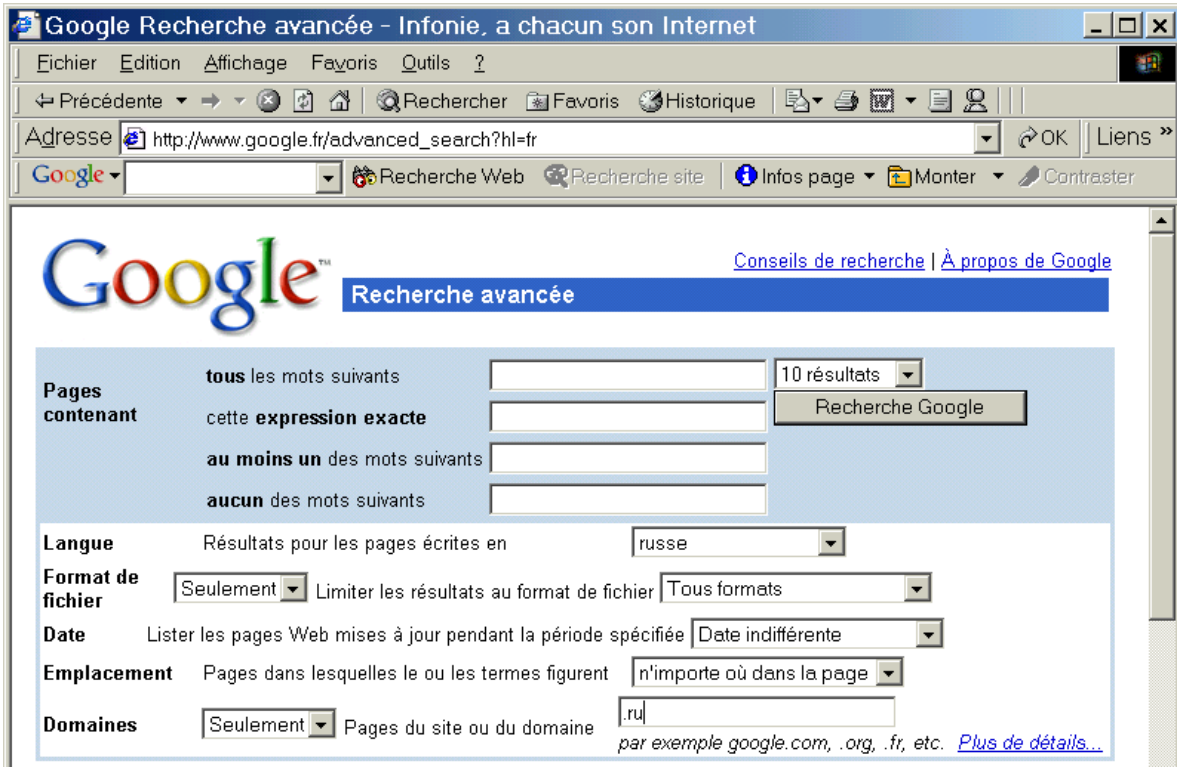
L'interface de Google propose une page « outils linguistiques » par l'intermédiaire de l'option « recherche avancée ». Celle-ci propose d'effectuer une requête :

- par langue. Une requête peut être effectuée en 35 langues dont le russe, grâce à un menu déroulant.

et

- par nom de domaine ou code domaine de pays (.ru, par exemple),

FIGURE N°2 : INTERFACE GOOGLE : RECHERCHE AVANCEE



De plus, Google offre la possibilité de visualiser les résultats « en cache », ce qui permet de repérer rapidement l'information recherchée dans une page très longue, où figure, par exemple, un texte de communication de congrès ou un article, sur un site dont la taille des caractères est très petite et non modifiable. Cette fonctionnalité permet également de consulter des pages qui ne sont plus visibles sur le site d'origine mais qui ont été « aspirées » par le moteur de recherche.

L'inconvénient majeur de ce moteur de recherche tient au fait qu'il classe les pages présentées comme résultats, notamment en fonction du nombre de sites qui pointent vers ces pages. Ce qui pénalise les pages récentes et celles qui sont liées à des domaines très pointus. Il se peut donc que des sites pertinents aient échappé à mes requêtes.

Mes premières recherches de centres de réflexion russes ont été effectuées, tout d'abord, à l'aide de requêtes en français via Google français, à l'aide de mots clés. La raison en est que la manipulation informatique[5] de bascule du clavier français au clavier russe nécessitait l'intervention de l'administrateur du lieu de stage, le seul à connaître le mot de passe indispensable. Ce contretemps a présenté l'intérêt de faire surgir des questions sur le fonctionnement de ce moteur de recherche.
Mes recherches se sont élargies ensuite, à partir de la page de Google russe, avec des mots-clés en russe.

[5] Cette manipulation est très simple à faire : cliquer sur Démarrer, choisir « paramètres », puis « panneau de configuration ». Cliquer sur l'icône « clavier », choisir l'onglet « langue ». Cliquer sur « Ajouter » et choisir la langue, en faisant défiler le menu déroulant. Cliquer sur OK. Le basculement d'un clavier sur l'autre se fait soit en cliquant sur le bouton « codage linguistique » en bas à droite de l'écran (petite icône carrée bleue), soit en effectuant une combinaison de touches à choisir. Une combinaison est prédéfinie par défaut.

Ne pas avoir disposé immédiatement du clavier russe m'a permis de constater qu'il est possible d'effectuer une requête sur le moteur de recherche Google, en français et d'obtenir une réponse en russe.

Cela signifie donc que :

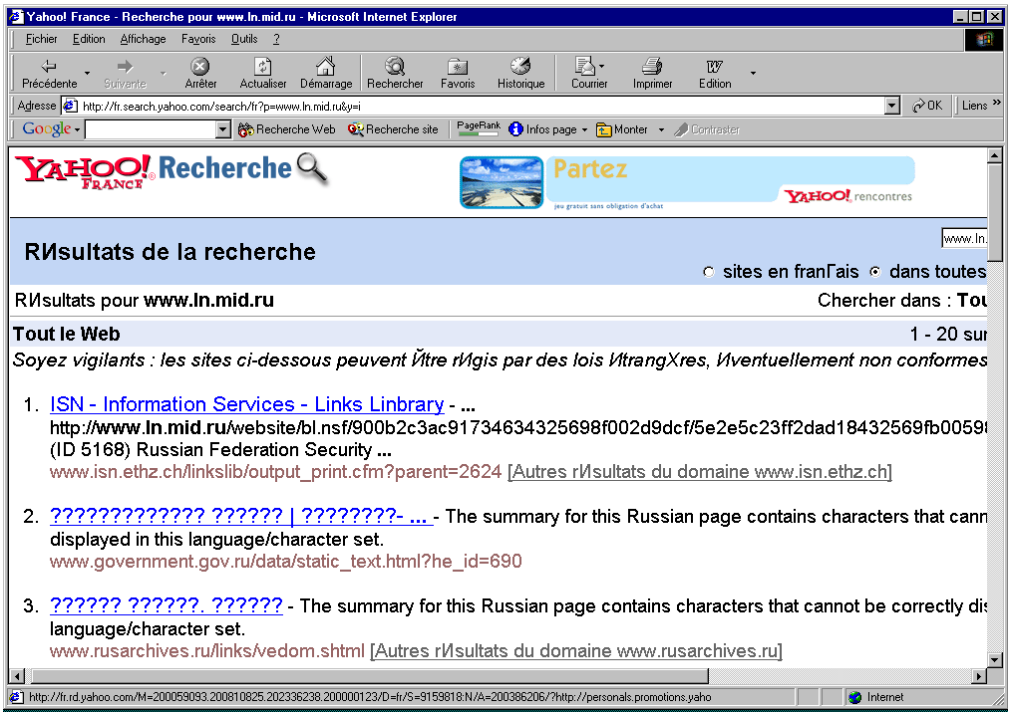
- l'affichage à l'écran de pages Web en caractères cyrilliques se fait sans problème avec le navigateur Microsoft Internet Explorer.
- La recherche bilingue franco-russe est possible via le moteur de recherche Google.

II.2. L'affichage à l'écran des caractères cyrilliques par des navigateurs, des annuaires et un moteur de recherche.

II.2.a / Le cas d'un annuaire de sites : Yahoo

Une tentative effectuée avec Yahoo ne donne pas de bons résultats en terme d'affichage. L'image copie d'écran ci-dessous montre que les caractères cyrilliques ne s'affichent pas du tout : certains liens menant à la page recherchée ne peuvent être affichés avec les caractères cyrilliques. Yahoo mentionne, du reste, à côté de ces liens le fait que le jeu de caractères n'est pas disponible. Bien que l'adresse du site recherché soit un « .ru », la police dont a besoin le navigateur pour afficher correctement les caractères est un jeu de caractères ukrainien.

FIGURE N° 3 : EXEMPLE D'UN PROBLEME D'AFFICHAGE DE CARACTERES NON LATINS PAR UN ANNUAIRE



II.2.b/ Affichage de pages Web

Il est arrivé, rarement, au cours de mes recherches, via Google, qu'un site russe ne s'affiche pas immédiatement avec les bons caractères. En effet, les problèmes d'affichage ne se posent pratiquement plus, ce que confirment deux auteurs pour qui « Sur Internet, auquel on accède couramment via un navigateur tel que Netscape ou Internet Explorer, les problèmes de décodage des pages ont pratiquement disparu » (40).

Lorsqu'un site ne s'affiche pas correctement en russe, il suffit de cliquer sur « Codage » dans l'onglet « Affichage », puis de sélectionner le bon parmi les quatre proposés, à savoir : Cyrillique (ISO), Cyrillique (KOI8-R), Cyrillique (Windows) et Cyrillique (KOI8-U) qui code les caractères pour l'ukrainien.

La présence d'une stagiaire spécialiste de la Chine, au centre documentaire du Campus de Défense, m'a permis de constater que si la question de l'affichage des caractères cyrilliques est vite résolue, l'affichage de caractères chinois paraît moins simple puisque l'ajout d'une police spécifique, téléchargée à partir d'un site chinois, a été nécessaire, le navigateur Internet Explorer n'en disposant pas.

Ces exemples conduisent à distinguer deux points. Le premier est le rendu visuel des caractères, c'est-à-dire l'affichage à l'écran qui vient d'être évoqué tandis que le second concerne l'analyse, le traitement, l'indexation et le stockage du contenu des pages par les moteurs de recherche.

III. Le codage des caractères non latins des pages de la Toile

L'absence de problèmes majeurs, (exceptée la question des polices disponibles ou pas), pour trouver des sites en caractères non latins s'explique par l'utilisation du codage Unicode qui permet de sortir du cadre étriqué des seuls caractères employés par l'anglais. Examinons tout d'abord le cas des adresses de sites qui restent, elles, encore prisonnières du codage des seuls caractères latins, mais peut-être plus pour très longtemps.

III.1. Les adresses électroniques de sites (URL)

En 1999, C. De Loupy notait que « les infrastructures et les normes utilisées sur la Toile portent encore largement la marque de son origine américaine » (18). Citons, à titre d'exemple, les Serveurs de Nom de Domaine (DNS) qui « traduisent » les adresses électroniques (URL) en adresse numérique comprise par une machine. Ces serveurs ne fonctionnent qu'en ASCII[6] (18), ce qui explique pourquoi les URL ne sont codées qu'à l'aide de l'ASCII et ne comportent donc que les lettres majuscules et minuscules de a à z, des chiffres de 0 à 9 et des tirets.
Mais cette situation pourrait changer puisque Microsoft permet d'élargir la prise en charge du jeu de caractères des URL au format Utf-8, grâce à Windows 2000 Server.
Les serveurs ayant intégré Windows 2000 Server restent compatibles avec les serveurs DNS utilisant l'ASCII.

[6] Cf. un développement succinct concernant l'ASCII en deuxième partie : II.1. L'ASCII, p 33.

Ceci est rendu possible notamment grâce l'utilisation des minuscules des noms codés au format Utf-8, avant la transmission des pages. Selon la RFC[7] 1035, tous les serveurs doivent être en mesure de comparer les noms sans tenir compte de la casse. Ainsi, ceux qui ne reconnaissent pas le format Utf-8 peuvent néanmoins recevoir et effectuer des comparaisons binaires des données. (19)

Par conséquent, une solution vient d'être apportée pour remédier à l'inconvénient de la limitation aux caractères latins non étendus dans les URL. En effet, cette limitation représente une perte de la valeur mnémonique de ces adresses pour toutes les personnes n'utilisant pas des caractères latins de façon courante !

III.2. HTTP, HTML et Unicode

III.2.a/ HTTP et MIME

Les échanges entre clients et serveurs sont gérés par un protocole HTTP (Hyper Text Transport Protocol), « en permettant aux premiers d'exprimer des requêtes et aux seconds de répondre en retournant les documents demandés » (35).
Pour les requêtes comme pour les réponses, HTTP utilise des en-têtes MIME (Multi Purpose Internet Mail Extension) pour échanger des méta-informations.
Dans une réponse, la ligne d'en-tête la plus importante est appelée Content-Type qui donne au client une information sur le type d'informations données par le serveur.
MIME constitue un ensemble de règles pour la mise en forme, la transmission et l'analyse de messages pouvant contenir n'importe quel type de données. Il encapsule les éléments, l'en-tête (qui contient le champ Content-Type) et le document (27).
MIME permet d'utiliser les alphabets codés sur plus de 8 bits, en rajoutant trois lignes dans l'en-tête de tout message échangé sur Internet. L'une d'entre elles indique le contenu du message, l'autre le codage utilisé et la troisième indique que MIME est employé.
Les serveurs électroniques n'intègrent pas tous des normes basées sur des codages à 8 bits. Ainsi il peut arriver qu'un texte codé sur 8 bits n'arrive pas à destination dans le même état que lors de son émission. La norme MIME permet de ramener un texte codé sur 8 bits à 7 bits.
Il constitue donc un outil important pour l'échange de textes contenant des jeux de caractères différents, y compris pour l'échange de courriers électroniques entre différents systèmes informatiques. Il autorise la composition et la consultation de messages avec d'autres jeux de caractères que celui de l'ASCII.

III.2.b/ HTML et Unicode

HTML (Hyper Hyper Text Markup Language) est un langage qui définit le contenu des pages Web et qui en permet l'affichage par les navigateurs.
Il est possible de visualiser le contenu d'une page HTML en regardant le code source. Pour afficher en caractères cyrilliques le code source qui s'écrit en HTML et donc est codé en Unicode, il suffit de cliquer sur « Affichage » dans la fenêtre du code source et de choisir une police cyrillique.

[7] RFC : Request For Comments : ils décrivent les standards pratiqués sur Internet

HTML a été conçu à partir de la norme de codage ISO 8859-1, c'est-à-dire le codage des alphabets d'Europe occidentale. Mais par la suite, HTML a été utilisé en d'autres langues avec d'autres jeux ou codages de caractères. Par défaut, la version HTML 4.0 utilise à présent Unicode et supporte particulièrement le format Utf-8. Grâce à ce format, toutes les formes d'écritures dans les pages HTML sont acceptées et affichées de façon implicite par les navigateurs.
En effet, le groupe de travail HTML de l'organisme chargé de créer et de maintenir à jour les standards Internet[8] a entrepris d'adopter Unicode comme jeu de référence. C'est le standard RFC 2070 « Internationalisation de HTML » qui définit cette adoption (34).
Cette internationalisation de HTML se fait par l'introduction d'éléments et d'attributs linguistiques, à savoir l'étiquetage linguistique (« Lang ») et la gestion de la bidirectionalité (« Dir ») qui n'intervient pas dans le cas de l'écriture cyrillique (38).

L'étiquetage linguistique consiste à identifier ces attributs linguistiques et des techniques de codage grâce à l'élément MÉTA. Il permet aux navigateurs de changer de fontes chaque fois qu'ils détectent ainsi un jeu de caractères codés différent.
Cette introduction d'éléments et d'attributs linguistiques constitue ce que E. Giguet et N. Lucas (40) appellent l'enrichissement des protocoles d'échange que représente MIME.

Pour élaborer des pages HTML multilingues, il convient :

- d'introduire le paramètre http « Charset » dans le champ « Content-Type » dans l'en-tête du document,
- de faire une déclaration « META » pour « http-equiv » avec « Content-Type » et une valeur appropriée pour « Charset »,
- de faire usage de l'attribut « Charset » au niveau des éléments qui font appel à des ressources externes.

III.3. La conversion en Unicode

En 1999, le moteur Alta Vista proposait un service de recherche de documents dans une langue spécifique (33). Ce service est possible grâce au stockage de l'information de pages écrites en différentes langues, dans un index unique, indépendamment des jeux de caractères utilisés pour l'écriture de cette information. Ce même moteur proposait également une recherche multilingue.

La technique employée consiste à coder toute les pages « aspirées » en Unicode qui permet le stockage de tous les caractères. La conversion des pages vers Unicode s'effectue en deux phases :

1/ identification du codage source de la page,
2/ transcodage du codage source vers Unicode.

La phase d'identification du codage source comprend plusieurs techniques qui sont les suivantes :

[8] Il s'agit de l'IETF (Internet Engineering Task Force) qui est membre du Consortium Unicode.

- celle d'Internet Explorer consiste à identifier le codage grâce au paramètre charset (= character set : jeu de caractères) dans la balise « Content-Type » d'un en-tête MIME définie par le protocole HTTP,

- identification grâce à l'attribut « Charset » dans une page HTML,

- celle de Mozilla de Netscape, par exemple, consiste à identifier les séquences de bits caractéristiques d'un codage ou la fréquence de séquence d'octets (40).

III.4. Le codage des pages en caractères cyrilliques

Les conditions nécessaires pour l'affichage des caractères cyrilliques sont les suivantes :

- le navigateur doit pouvoir comprendre la norme utilisée pour le codage des caractères,
- utilisation de la balise

CONTENT=″text/html ; charset=ISO-8859-5″

ou

CONTENT=″text/html ; charset=Windows1251

ou

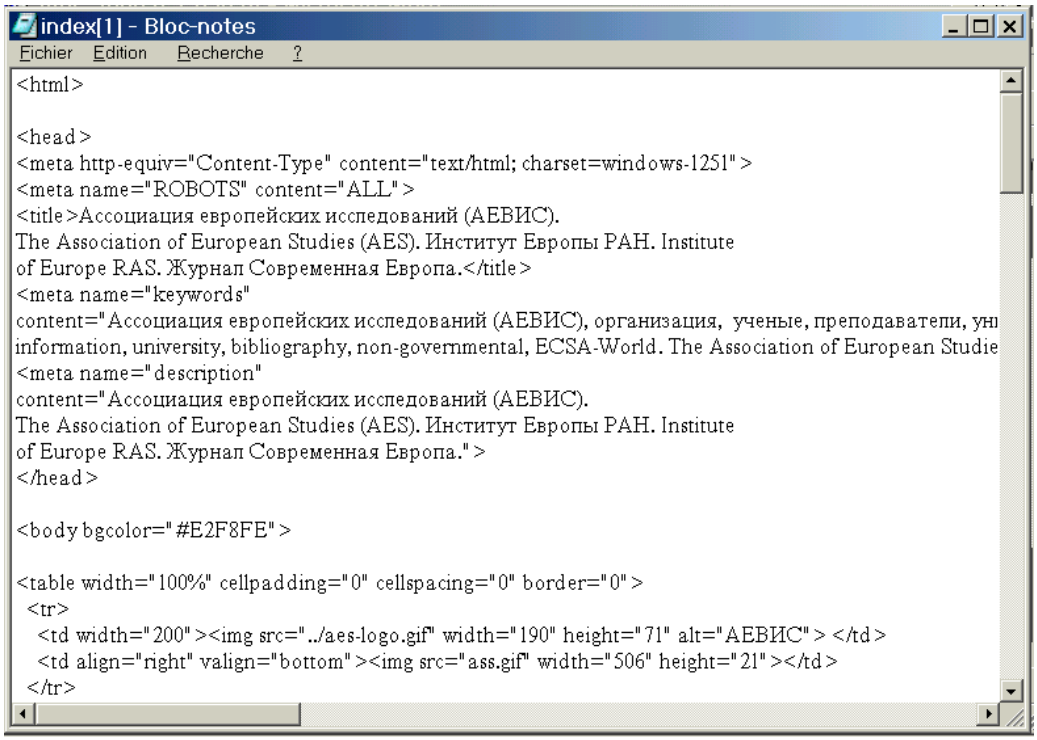
CONTENT=″text/html ; charset=KOI8-R

D'après la page intitulée "L'écriture cyrillique" du site www.culture.fr/culture/edm/fr/Cy/cy03.html, « le codage le plus utilisé sur les clients http est le KOI-8 (environ 50% des sites en cyrillique). » Mais le codage des caractères cyrilliques le plus employé par les créateurs de sites russes, recensés dans le cadre de ma mission de stage, semble être plutôt Windows 1251[9].

La copie d'écran du code source du site de l'association des études européennes montre que cette consigne a été respectée.

[9] Un test, effectué sur 39 sites parmi les 118 recensés dans le cadre de ma mission, donne un résultat de 29 codes sources indiquant un jeu de caractères Windows 1251.

FIGURE N°4 : LE CODE SOURCE D'UN SITE RUSSE AVEC LA BALISE META INDIQUANT LE JEU DE CARACTERES WINDOWS 1251



La navigation de site en site de Russie, d'Asie centrale et d'Europe centrale ayant abouti à la collecte d'adresses de sites de centres de réflexion, de noms de périodiques publiés par certains de ces centres, ainsi que plus de quatre-vingts noms de spécialistes des questions de défense travaillant et résidant dans l'espace ex-soviétique et en Europe centrale, il a fallu ensuite les saisir dans une base de données, dont les champs avaient été définis préalablement à ma mission.

Cette opération a soulevé des questions relatives à la romanisation des noms des chercheurs et des collectivités auxquelles ils appartiennent et à la forme courante à adopter.

IV. Alimentation de la base "Annuaire des chercheurs de défense" (logiciel Access)

La première difficulté est donc apparue lorsqu'il s'est agi de rentrer les premiers noms dans l' « Annuaire des chercheurs de défense ». Les questions de langue et de transcription en caractères romains se sont alors posées.
Précisons, tout d'abord, les termes romanisation, translittération, transcription.

IV.1. Romanisation, translittération, transcription,

L'AFNOR donne les définitions de ces termes (32) :

Romanisation
La romanisation est une conversion d'écriture non latine dans l'alphabet latin.

L'AFNOR admet qu'il existe plusieurs niveaux de conversion. Ils sont au nombre de trois. Le premier est la translittération rigoureuse, le second la conversion simplifiée, le troisième la conversion populaire.

- La translittération rigoureuse est complètement réversible car elle se fait caractère par caractère. Elle n'admet pas de variantes.
- La conversion simplifiée autorise des variations nationales ou régionales, notamment pour pouvoir utiliser des machines qui n'acceptent pas tous les caractères de l'alphabet exigés pour la conversion rigoureuse. Elle peut faire l'objet de normes et ou d'accords internationaux.
- La conversion populaire tient compte obligatoirement d'habitudes phonétiques ou graphiques et ne peut être que nationale.

Translittération

« 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 ».

Le terme « caractère » désigne, selon l'AFNOR, « un élément d'un système d'écriture, alphabétique ou non, représentant graphiquement un phonème[10], une syllabe, un mot, voire un trait prosodique d'une langue ».

Transcription

« 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 translittération est une opération rigoureuse de conversion caractère par caractère d'une langue à l'autre et permettant la réversibilité complète. Elle diffère de la transcription, qui cherche simplement à fournir un équivalent dans l'alphabet de conversion en utilisant les lettres permettant de rendre les sons de la langue convertie » (Marc-André Roberge, 31).

IV.2. Romanisation et base de données : cas du logiciel Access

Des experts de la BnF affirment "La technique rend possible la gestion de notices en plusieurs langues et en plusieurs écritures". (E. Freyre, F. Bourdon, 13). Cela se vérifie lorsque les logiciels ont intégré le format de codage Unicode. Lorsque ce n'est pas le cas, la gestion de caractères non latins est complexe, voire impossible comme pour le logiciel Access 97 SR-1 qui ne code pas du tout les caractères cyrilliques ; la simple saisie de ces caractères est impossible, à moins d'installer la prise en charge d'Unicode sous Windows[11].

Les caractères cyrilliques n'étant pas pris en compte par le logiciel Access 97 SR-1 sur mon lieu de stage et, par ailleurs, le travail de recherche de ces noms d'experts s'étant fait à partir de documents en anglais ou dans diverses langues d'Europe centrale et orientale, il a fallu choisir la ou les écriture(s) à employer ainsi que la

[10] Phonème : Élément sonore du langage articulé, considéré du point de vue physiologique et acoustique (Le Petit Robert, 1989)

[11] Cf. « Comment installer la prise en charge d'Unicode sous Windows, annexe 1 p. 78.

norme de romanisation. L'idée, qui semblait la plus simple, a priori, consistait à adopter la norme NF ISO 9 (juin 1995) de translittération[12].

Mais, la norme française de translittération ISO 9/95 a été rapidement éliminée, en raison, notamment, des problèmes posés par la saisie de quatre signes diacritiques (Č, Ž, Š, Ŝ) nécessaires à sa mise en application. Le clavier français ne permet en effet pas de les saisir aisément, sauf à passer par la table des caractères Unicode, proposée en accessoire par les logiciels Access 97 SR-1 et Word 97 SR 2 ou à personnaliser le clavier[13]. Cette manipulation est peu ergonomique. (L'autre solution aurait consisté à implanter Unicode dans Windows 97 et dans Access.)

La solution de translittération selon la norme ISO étant écartée, restait celle d'une transcription française ou anglo-saxonne. C'est cette dernière qui fut retenue par le Centre Documentaire du Campus de Défense. Ce choix était basé sur l'hypothèse selon laquelle la plupart des utilisateurs de la base connaissent davantage cette forme que les autres. Mais ce type de transcription présente également des inconvénients, puisque qu'il n'est pas strictement réversible et que le passage de cette forme à une translittération du type ISO ou le retour à la forme cyrillique peut être hasardeux.

Ces questions m'ont amenée à examiner la situation qui prévaut dans les catalogues en ligne de la BnF, de l'ABES[14] et de la FNSP.

V. Catalogage et translittération

Nous avons choisi d'interroger des catalogues en ligne avec deux noms d'auteurs connus de la littérature russe et un titre de périodique. Nous avions ainsi de très grandes chances de les trouver. De plus, leur orthographe en russe contient des lettres non latines dont la romanisation pose des problèmes. Mais cette recherche n'invalide pas le raisonnement qui vaut aussi pour des noms de chercheurs dont le renom n'est pas le même que celui d'écrivains célèbres. Elle met en lumière la question de la forme courante pour n'importe quel nom, adoptée par l'une ou l'autre des agences bibliographiques.

L'exemple de Tchekov, (Чехов) constitue une bonne illustration du problème avec ses deux lettres Ч et Х. La forme Tchekov n'est, du reste, qu'une forme actuelle parmi d'autres, comme Tchékov, Tchékhov, Tchekhov (sans accent)... Ces formes consacrées par l'usage, ne correspondent pas à la translittération selon la norme ISO 9/95 qui est Čekov, dite aussi « forme savante ». Mais ces formes étant connues, il n'est pas question de les ignorer, puisque tout usager peut employer l'une ou l'autre lors d'une requête, sans même avoir à l'esprit l'existence des autres formes.

Les trois premières lettres en français (tch) correspondent à la lettre Ч en russe. Ces trois lettres représentent donc une forme transcrite en français de la lettre russe. Mais elles ne correspondent pas à la forme translittérée rigoureuse selon la

[12] Cf. la table en annexe 2 p. 80.
[13] Voir les manipulations possibles pour saisir les quatre caractères diacrités en annexe 3 p. 81.
[14] ABES : Agence Bibliographique de l'Enseignement Supérieur

norme ISO 9/95 qui est Č. L'autre lettre russe X est transcrite tantôt par un kh, tantôt par un h qui suit la norme de translittération ISO.
L'autre exemple de nom d'auteur est Saltykov-Chtchedrine (Салтыков-Щедрин), nom composé et dont le Щ est lui aussi source de problème en matière de romanisation.

Ces recherches ont été effectuées selon plusieurs graphies dans deux catalogues en ligne, le SUDOC et BN-Opale Plus[15] et à partir d'ordinateurs PC et Macintosh. Les résultats sont examinés ci-dessous.

V.1. La romanisation des caractères cyrilliques dans le SUDOC

L'examen de quelques notices dans le catalogue du SUDOC est surprenant de prime abord, puisque une grande hétérogénéité y règne en matière de romanisation. Il peut être également surprenant pour l'utilisateur final qui n'est pas nécessairement familier de ces questions et qui ne s'intéresse a priori qu'aux réponses obtenues.

V.1.a/ Recherche effectuée avec "Tchékov"

Le nombre de résultats est 21. Sur les dix premiers résultats, apparaissent quatre transcriptions, dans le titre d'une œuvre ou comme auteur : Tchékov, Tchekov, Tchekhov, et Cekhov.

V.1.b/ Recherche effectuée avec "Tchékhov"

Le nombre de résultats a varié, pour des raisons inconnues, de 729 à 172 entre deux requêtes similaires, effectuées à 24 heures d'intervalle.
Parmi les dix premiers résultats figurent deux graphies : Tchekhov et celle de la requête, Tchékhov.
En ouvrant une des notices bibliographiques, on trouve d'autres graphies, comme Cekhov.

V.1.c/ Recherche effectuée avec "Tchekov"

La même requête effectuée à très peu de temps d'intervalle donne une fois 521 et une autre 21 résultats.
Tous ces résultats confondus donnent les formes : Tchékhov, Tchekhov, Cehov (sans signe diacritique), Tchekov, Tchékov.

V.1.d/ Recherche effectuée avec "Tchekhov"

172 résultat sont annoncés. Deux graphies sont représentées : Tchékhov et Tchekhov.

V.1.e/ Recherche effectuée avec Cekov (sans signe diacritique)

Le moteur de recherche indique qu'il n'a pas trouvé de résultat.

V.1.f/ Recherche effectuée avec Čekov

Pour saisir la lettre Č, le recours à la table des caractères Unicode ou aux caractères spéciaux est impossible à partir de l'interface du navigateur. Alors, une « solution de secours » consiste à copier la lettre à partir d'un fichier Word et à la

[15] BN-Opale Plus : catalogue en ligne de la BnF

coller dans la fenêtre de l'interface du moteur de recherche du SUDOC. Cette tentative échoue puisque la lettre copiée n'est pas conservée (elle est transformée en une petite barre verticale) et ignorée par le moteur de recherche.

Récapitulation des résultats des recherches effectuées dans le catalogue du SUDOC Interface Web à partir d'un ordinateur PC

Requête	Tchékov	Tchékhov	Tchekov	Tchekhov	Cekov	Čekov
Nombre de Réponses	21	729 ou 172	521 ou 21	172	0	Requête impossible
Transcriptions obtenues dans les 10 premiers résultats	Tchékov, Tchekov, Tchekhov, Cekhov	Tchékhov Tchekhov	Tchékhov Tchekhov Tchekov Tchékov Cehov	Tchékhov Tchekhov		

V.1.g/ Recherche effectuée avec Saltykov-Chtchedrine

Une autre recherche effectuée, toujours dans le SUDOC, sur le nom de Saltykov-Chtchedrine donne des résultats non moins disparates.
Tout d'abord, l'interrogation avec le nom composé en entier et la graphie « Saltykov-Chtchedrine » ne donne aucune réponse. Il faut réduire la requête à la première partie du nom "Saltykov". Ce qui est curieux puisqu'en principe au moins une zone de la notice d'autorité contient les deux parties d'un nom composé, ce qui implique que le résultat avec le nom entier apparaît nécessairement.

Parmi les 10 premiers résultats sur 60, pas moins de trois graphies s'affichent pour la deuxième partie du nom, mais pas une seule n'est rigoureusement conforme à la norme ISO. Il s'agit de :

- ^Sedrin : le signe diacritique est placé avant le S,
- Sedrin : le signe diacritique manque pour se conformer complètement à la norme ISO,
- Shchedrin : il s'agit d'une transcription anglo-saxonne.

V. 2. La translittération des caractères cyrilliques dans le catalogue BN-Opale Plus de la BnF

Les mêmes requêtes ont été effectuées dans le catalogue BN-Opale Plus, via une interface Web.
La présentation des réponses par le moteur de BN-Opale Plus n'est pas identique à celle du SUDOC. Le catalogue BN-Opale Plus propose des résultats sous forme d'entrées ou de notices abrégées. Par entrées, il faut comprendre une liste de noms dont les premières lettres sont celles du nom qui a servi à la requête. Lorsque le moteur de recherche ne trouve qu'un seul nom, les notices abrégées sont affichées immédiatement. Lorsqu'il trouve plusieurs noms dont les premières lettres correspondent à la requête, il les affiche. Les notices sont visibles après un clic sur le nom recherché.

V.2.a / Requête effectuée avec la graphie Tchékov

Les entrées obtenues sont : Tchekoff, Tchékov, avec un renvoi à Tchekhov, Tchekov, avec un renvoi à Cehov. On obtient aussi d'autres noms commençant par "Tchekh" et "Tchék".

La mention d'un renvoi, faite dans les résultats de cette première requête, montre que Tchékov et Tchekov sont des formes rejetées et que des notices d'autorité sont liées à des notices bibliographiques.
Notons que le renvoi vers la forme Cehov pose un problème. D'une part, un renvoi est, par principe, fait vers une forme retenue et d'autre part, la forme Cehov n'est justement pas une forme normalisée (puisqu'il lui manque un signe diacritique) et donc ne peut pas être considérée a priori comme une forme retenue.

V.2.b / Requête effectuée avec la graphie Tchékhov

Les graphies données sont Tchekhov, Tchékhov. Ces résultats sont identiques à ceux de la requête effectuée avec Tchekhov.

V.2.c / Requête effectuée avec la graphie Tchekov

Les graphies données sont : Tchékov, Tchekov.

V.2.d/ Requête effectuée avec la graphie Tchekhov

Les résultats sont identiques à ceux de la requête effectuée avec Tchékov.

V.2.e/ Requête effectuée avec Cehov

Parmi des résultats sans intérêt figurent : Cehov et Čehov avec un renvoi à Tchekhov. Nous remarquons que la forme Čehov qui pourrait être considérée comme la forme retenue (puisque totalement conforme à la forme ISO) ne l'est pas.
On note à ce propos que le caractère Č est bien pris en compte par le moteur de recherche. Il apparaît dans les résultats mais ne peut être manifestement pas utilisé à l'interrogation du catalogue en ligne.

V. 2.f/ Requête effectuée avec Čehov

Le Č, « saisi » dans l'interface du moteur de recherche de BN-Opale Plus par un copier-coller effectué à partir d'un fichier Word, est pris en compte par ce moteur, mais interprété comme un È et les résultats sont très éloignés de ce qui était escompté.

Récapitulation des résultats des recherches effectuées dans BN-Opale Plus
Interface Web à partir d'un ordinateur PC

Requête	Tchékov	Tchékhov	Tchekov	Tchekhov	Cekov	Čekov
Transcriptions obtenues parmi les résultats	- Tchekoff, - Tchékov, avec un renvoi à Tchekhov - Tchekov, avec un renvoi à Cehov - Plusieurs noms commençant par "Tchekh*", "Tchék*",	-Tchekhov -Tchékhov	Mêmes résultats qu'avec Tchékov	Mêmes résultats qu'avec Tchékhov	- Cekov - Čekov avec un renvoi à Tchekhov	Pas de résultats

V.2.g/ Recherche effectuée avec Saltykov-Chtchedrine

Cette recherche effectuée, via une interface Web, à partir d'un ordinateur PC, avec le nom composé entier, donne immédiatement 40 réponses sous formes de notices. Les titres des ouvrages sont translittérés selon la norme ISO 9. Les signes diacrités sont affichés. De petits carrés s'affichent également entre deux lettres.
L'impression des résultats dans les locaux de l'INTD ne rend pas les quatre signes diacrités à savoir Ŝ, Š, Č, Ž, mais elle les rend sur une autre imprimante. Donc le rendu à l'impression varie en fonction de la configuration de l'imprimante.

V.3. Cas de la FNSP

À la lumière de la consultation en ligne, à partir d'un ordinateur PC, de la liste des périodiques analysés par la bibliothèque de la FNSP, il apparaît que la romanisation de "Российский Экономический Журнал" semble bien suivre la norme de translittération ISO 9/95 aux signes diacritiques (ž et č) près. On obtient à l'écran : Rossijskij èkonomiceskij zurnal, au lieu de Rossijskij èkonomičeskij žurnal.

V. 4. Conclusions

Tous ces résultats permettent de faire plusieurs remarques. L'une d'elles consiste à mettre en lumière le fait que les problèmes de romanisation et de translittération d'une part et d'informatique (codage et / ou saisie et / ou affichage) d'autre part apparaissent comme mélangés et il convient de les distinguer.

V. 4.a/ Multiplicité de systèmes de romanisation au sein d'un même catalogue ou d'une même notice.

Les notices du SUDOC ont des provenances diverses, dont OCLC[16] et SF[17], et elles ont des procédés de romanisation différents. Une même notice SUDOC peut contenir des accès relevant de plusieurs systèmes de translittération. À titre d'exemple, citons le résultat 9 de la requête Saltykov dans SUDOC qui est

[16] OCLC : Online Computer Library Center
[17] SF : Sibil France

une notice SF. Le titre de la notice contient la graphie Saltykov-Sedrin (translittération ISO sans signe diacrité), tandis que la transcription française (Saltykov-Chtchedrine) figure dans l'accès « Sujets ».

Parmi les explications, mentionnons notamment le fait que lors de la rétroconversion des notices, toutes n'ont pas été chaînées à des notices d'autorité (et en particulier celles qui relevaient d'une transcription anglo-saxonne), et que les notices d'autorité auteur et matière n'ont pas été fusionnées.
Ces problèmes de chaînage entre notices résolus permettraient d'obtenir des résultats plus cohérents, présentant une ou deux formes retenues.

La situation qui prévaut est ainsi résumée par D. Duclos-Faure (11), "Aujourd'hui tous les experts soulignent la grande hétérogénéité des accès auteurs dans SUDOC au sein parfois d'une même notice, hétérogénéité résultant d'une fusion de notices de diverses sources ayant des systèmes de romanisation différents (ISO R9/68, ISO 9/86/95, formes anglo-saxonnes issues d'OCLC et formes courantes françaises issues des fichiers de la BnF) et due au fait que peu d'auteurs d'écritures non latines sont aujourd'hui présents dans les fichiers d'autorité à la BnF. En effet, les ouvrages en caractères non latins ne sont catalogués et translittérés sur la base BN OPALE que depuis 1997 et hormis les périodiques, les fonds en caractères non-latins de la BnF ne sont pas rétroconvertis".

V.4.b/ Le cas de la BnF

Il existe des notices d'autorité dans BN-Opale Plus qui réunissent les différentes formes de romanisation et les répartissent en formes retenue et rejetée. Elles permettent ainsi de présenter dans un même lot de résultats un certain nombre de notices, quelle que soit la forme employée dans la requête.
Pour les noms étrangers, deux formes sont retenues dans ces notices d'autorité, à savoir la forme courante et la forme savante. La BnF privilégie une forme courante, choisie parmi d'autres en fonction de divers critères, dont la fréquence d'usage.

Nous avons cherché une notice d'autorité pour l'auteur Tchékov, via l'interface Web, sur un ordinateur PC et non pas via l'accès en salle de lecture, dans les locaux de la BnF. Pour faire apparaître une telle notice, il suffit de cliquer sur l'entrée qui correspond au nom de l'auteur recherché. Il s'agit d'un lien lorsqu'elle est de couleur rouge bordeaux. Les entrées « Tchekhov » et « Cehov » sont de tels liens. Nous avons donc obtenu deux notices d'autorité, présentées ci-dessous.

FIGURE N°5 : NOTICE D'AUTORITE PERSONNE PHYSIQUE : TCHEKHOV (BN-OPALE PLUS)
Affichage sur un ordinateur PC

Tchekhov, Anton Pavlovitch (1860-1904) forme courante autre système de translittération à valeur internationale
Cehov, Anton Pavlovic forme savante à valeur internationale système ISO

Nationalité : Russie, Fédération de
Langue : russe
Sexe :
Responsabilité : Auteur
Naissance : 1860 -
Mort : 1904
La vedette peut être employée dans une vedette RAMEAU
La vedette ne peut s'employer qu'en tête de vedette
La vedette utilisée en zone 6XX n'admet pas de subdivision géographique
Forme(s) rejetée(s) :< Tchékov, A. P.
Sources : La Cerisaie / Anton P. Tchekhov ; trad. de Jean-Claude Carrière ; conseillère pour la langue russe Lusia Lavrova, 1988.
- GDEL
- BN Service russe
Notice n° : FRBNF11926133

Dans la première notice ci-dessus (figure N°5), la première forme courante retenue est Tchekhov. La seconde forme retenue est la forme savante, Cekov, qui figure ici sans le signe diacritique, ce qui constitue une anomalie, puisque ces signes sont, en principe, rendus à l'affichage. La seule forme rejetée est Tchékov, ce qui est également surprenant puisque d'autres graphies existent et qu'elles figurent dans les entrées du catalogue. Cette notice est un produit de la fusion qui s'est opérée entre fichiers d'autorité auteurs et fichiers d'autorité matière à la BnF.

FIGURE N°6 : NOTICE D'AUTORITE PERSONNE PHYSIQUE : CEHOV (BN-OPALE PLUS)
Affichage sur un ordinateur PC

Cehov, Anton Pavlovic (1860-1904) *forme savante à valeur internationale système ISO russe*
Tchekhov, Anton Pavlovitch (1860-1904) *forme courante romanisation sans système de translittération connu russe*

Nationalité : Russie, Fédération de
Langue :
Sexe : Masculin
Responsabilité : Auteur
Naissance : 1860-01-17 -
Mort : 1904-07-02
La vedette ne peut pas être employée dans une vedette RAMEAU

Forme(s) rejetée(s) :
< Chekov, Anton *russe*
< Tchékov, A. P. *russe*
Sources : BN-APP (Opale), 1996-07-17. - Nathan technique 879141, Lectures suivies

Notice n° : FRBNF13967344

La seconde notice d'autorité (figure N°6) présente les mêmes formes courante et savante, mais dans l'ordre inverse, avec un commentaire (en italique), pour la forme courante, différent de celui de la précédente notice : *forme courante romanisation sans système de translittération connu russe / forme courante autre système de translittération à valeur internationale.*
Les formes rejetées sont, elles, différentes : la première notice d'autorité ne fait figurer que la forme Tchékov, tandis que la seconde fait aussi apparaître la forme Chekov. En revanche, les autres formes Tchekhov et Tchékhov sont toujours ignorées en forme rejetée. Cette notice n'est pas conforme et sera probablement éliminée.

Par définition, une notice d'autorité doit être unique pour un auteur. Toutes les notices bibliographiques et notices d'autorité n'ont donc pas été « nettoyées » et ont toutes été conservées dans le catalogue. Elles sont éliminées au cours du temps.

Les résultats d'une requête dans le catalogue BN-Opale, effectuée toujours via le même navigateur, mais sur un ordinateur Macintosh, apporte un éclairage quant à la question de l'affichage des signes diacritiques, puisque sur la figure N°7, ci-dessous, le signe diacritique apparaît, mais décalé par rapport à son emplacement théorique.

FIGURE N°7 : NOTICE D'AUTORITE PERSONNE PHYSIQUE :TCHEKHOV (BN-OPALE)
Affichage sur un ordinateur Macintosh

Tchekhov, Anton Pavlovitch (1860-1904) *forme courante autre système de translittération à valeur internationale*
C˘ehov, Anton Pavlovic˘ *forme savante à valeur internationale système ISO*
Nationalité : Russie, Fédération de
Langue : russe
Sexe :
Responsabilité : Auteur
Naissance : 1860 -
Mort : 1904
La vedette peut être employée dans une vedette RAMEAU
La vedette ne peut s'employer qu'en tête de vedette
La vedette utilisée en zone 6XX n'admet pas de subdivision géographique
Forme(s) rejetée(s) :
< Tchékov, A. P.
Sources : La Cerisaie / Anton P. Tchekhov ; trad. de Jean-Claude Carrière ; conseillère pour la langue russe Lusia Lavrova, 1988
. - GDEL
. - BN Service russe
Notice n° : FRBNF11926133

V. 4.c/ Multiplicité de systèmes de romanisation adoptés par les différents catalogues.

La norme de translittération ISO 9/95 est en principe adoptée par la BIULO[18], la BnF, la Sorbonne, la BDIC[19], la FNSP.
Les fichiers manuels de la BIULO, BDIC, BNF et du Centre d'Études slaves ne contiennent, eux, que des notices en caractères originaux.

[18] BIULO : Bibliothèque Internuniversitaire des Langues Orientales
[19] BDIC : Bibliothèque de Documentation Internationale Contemporaine

Translittération	Fichier informatisé	Fichier manuel
Bibliothèque de la Sorbonne	ISO 9/95 dans SIBIL	Caractères originaux
BIULO	ALA-LC dans OCLC	Caractères originaux
BDIC	Caractères originaux dans Aleph 5000	Caractères originaux
BnF	ISO 9/95 dans BN-Opale Plus	Caractères originaux
FNSP	ISO 9/95	-
Centre d'études slaves	-	Caractères originaux

Source : Catalogage des caractères non-latins et latins étendus. Groupe d'experts pour le cyrillique. Synthèse des travaux. Janvier 2002

V. 4.d/ Aspects informatiques

Nous avons pu constater que la norme ISO 9/95 est appliquée, mais que l'affichage des signes diacritiques est problématique.

Les signes diacritiques sont affichés dans les notices d'autorité du catalogue BN-Opale Plus, lorsque celles-ci sont consultées à partir des postes dans les locaux de la BnF. Leur affichage n'est pas systématique lorsqu'une interface Web est placée entre l'utilisateur et le moteur de recherche de la BnF, comme le montrent les notices d'autorité de Tchekov, affichées sur un ordinateur PC ou Macintosh.

Par ailleurs, le type de résultats obtenus suite à l'interrogation dans les deux catalogues avec la graphie Čekov permet de constater que les deux moteurs de recherche des catalogues SUDOC et BN-Opale Plus ne fonctionnent pas de la même façon vis-à-vis de la saisie des signes diacritiques. Les exemples de requêtes effectuées ci-dessus montrent la lourdeur qu'implique l'introduction de ces différentes formes, dans un catalogue informatisé. Si elles ont été abandonnées à l'interrogation dans le cas de la BnF qui a fait le choix de ne pas utiliser les signes diacritiques à l'interrogation mais de les rendre à l'affichage, dans ses salles de lecture, c'est aussi peut-être pour simplifier l'interrogation en ne n'alourdissant pas avec l'utilisation de clavier virtuel, par exemple.

La situation présentée ci-dessus n'est pas exceptionnelle, comme l'explique M. Ben Henda (45) : « Les techniques de codages des ressources bibliographiques ont toujours été disparates, non uniformes et non exhaustives : ce qui a toujours engendré l'exclusion de certains fonds linguistiquement "complexes" du fonds général de certaines bibliothèques. Les bibliothèques et les centres de documentation ont cependant besoin d'un outillage linguistique simultané et multiscript. L'absence d'un consensus universel pour l'adoption d'un jeu de caractères unifié a fait que chaque structure se débrouillait pour adopter ses propres méthodes de traitement multilingues. »

Le consensus universel évoqué par M. Ben Henda étant apparu et en cours d'implantation, l'importance qu'il va prendre dans le monde de la documentation nous conduit à étudier les grandes lignes de sa conception.

Deuxième Partie : Unicode et Caractères cyrilliques

« L'écriture est depuis toujours étroitement liée aux nombres » (24)

Dans un premier temps, sera exposée la place de la langue russe dans la famille des langues slaves et l'utilisation des caractères cyrilliques par diverses langues. Puis nous examinerons la question des différents codages informatiques des caractères latins et non latins, l'avènement du système de codage universel Unicode et ses grands principes de fonctionnement, puis la place particulière qu'y tiennent les caractères cyrilliques.

I. La langue russe dans la famille des langues slaves et les caractères cyrilliques.

La langue russe appartient à la famille des langues slaves et elle est alphabétique. Elle s'écrit à l'aide de caractères cyrilliques, de gauche à droite et elle est bicamérale, ce qui signifie qu'elle distingue les lettres majuscules des minuscules.
Ceci en fait un cas plus simple que les langues syllabiques, idéographiques, idéographique-syllabiques (japonais), ou idéographiques et alpha-syllabiques par composition (coréen), du point de vue du codage informatique.
Toutes les écritures n'empruntent pas le même sens. Celui-ci peut varier dans un plan : il peut être de haut en bas, de gauche à droite, de droite à gauche, il peut aussi alterner de gauche à droite puis de droite à gauche à l'image des sillons d'un labour. Ce sens peut aussi être cruciverbé. Cette diversité de direction des écritures posent évidemment des problèmes de stockage et de placement des caractères dans une mise en page informatique (H. Hudrisier, 43). Nous n'aborderons donc pas ces aspects complexes puisque l'écriture en caractères cyrilliques ne présente pas ces spécificités.

Toutes les langues utilisant les caractères cyrilliques ne sont pas slaves. En effet, si ceux-ci sont utilisés par plus de soixante-dix langues, dont le russe, l'ukrainien, le biélorusse, le bulgare, qui sont des langues slaves, certaines langues du groupe turco-altaïque comme l'azéri, le kazakh, le kirghize les emploient aussi.
Ce chiffre varie dans le temps, suivant les politiques linguistiques adoptées par les différents régimes et les réévaluations idéologiques des pays concernés.
Ainsi, la langue ouzbèque, qui n'est pas une langue slave et qui a connu plusieurs alphabets au cours du XXème siècle (simplification de l'alphabet arabo-persan puis latinisation en 1926, cyrillisation entre 1936 et 1940), est de nouveau en train de passer des caractères cyrilliques aux caractères latins, suivant ainsi la décision des autorités ouzbèques prise en octobre 1993, deux ans après l'indépendance de l'Ouzbékistan.
En matière de translittération des caractères cyrilliques, employés par des langues non slaves, la norme ISO 9/95 ne serait pas d'un usage très répandu, selon le groupe d'experts pour le cyrillique du CLENOL[20] (5). Les systèmes privilégiés sont ceux qui rendent compte d'une langue plutôt que d'un alphabet. Cet état de fait ne présente sans doute pas d'inquiétudes, tant que les fonds et les échanges avec les

[20] Groupe de travail sur le catalogage des caractères non latins et latins, mis en oeuvre par la sous-direction des bibliothèques et de la documentation du ministère de l'Éducation nationale.

pays concernés restent faibles, mais, il posera des problèmes à terme, si ces échanges venaient à s'intensifier.
Il est à noter, par ailleurs, que si le russe est la langue la plus parlée dans la Fédération de Russie, celle-ci compte au total douze langues en usage.

Le tableau ci-dessous présente une liste non exhaustive des langues à écritures multiples, parlées dans l'ex-URSS.

FIGURE N°8 : LANGUES A ECRITURES MULTIPLES DE LA COMMUNAUTE DES ÉTATS INDEPENDANTS ET DES BALKANS

(Cette liste est non-limitative. L'usage de plusieurs écritures peut être simultané ou successif dans le temps.)

langue	écriture	nom particulier	pays	remarque
serbo-croate	latin	croato-serbe, croate	Croatie actuelle, en partie Fédération de Yougoslavie actuelle (au Monténégro) ; ancienne Yougoslavie (avant éclatement)	✓ Slaves du Sud catholiques ✓ le croate a aussi été écrit dans une écriture particulière, le *glagolitique*, utilisée jusqu'au XIX^{e} siècle dans des ouvrages liturgiques
	cyrillique	serbo-croate, serbe	actuelle Fédération de Yougoslavie (Serbie, en partie Monténégro) ; ancienne Yougoslavie (avant éclatement)	✓ Slaves du Sud orthodoxes
roumain	latin		Roumanie, Moldavie depuis quelques années	✓ l'écriture cyrillique a été utilisée jusqu'au XIX^{e}
	cyrillique étendu, puis latin	moldave	Moldavie	✓ l'écriture latine a été réintroduite depuis 1989
azéri	arabe étendu puis latin étendu, puis cyrillique étendu, puis latin étendu	azéri azerbaïdjanais	République d'Azerbaïdjan (ancienne république soviétique)	✓ quel degré de différence avec l'azéri du sud ?
kurde	arménien, puis latin étendu, puis cyrillique étendu	kurmandji ?	Arménie (ex-RSS d'Arménie)	✓ encore en cyrillique actuellement ?
persan oriental	arabe étendu, puis latin étendu, puis cyrillique étendu, puis tentative de retour à l'arabe étendu (depuis 1989)	tadjik	République du Tadjikistan (ancienne république soviétique)	✓ forme de persan oriental dont les différences dialectales ont été volontairement accentuées, mais qui sont en partie moins « visibles » lorsque l'écriture arabe est utilisée
langues de peuples musulmans de l'ex-URSS	arabe étendu	kirghize, ouzbek, tatar, ouïghour, certaines langues du Caucase….		
	latine étendu			à partir de 1920-1930
	cyrillique étendu			à partir de 1930-1943
	arabe étendu/ latin étendu			depuis 1989 environ

Source : Extrait du tableau LANGUES À ÉCRITURES MULTIPLES de la Communauté des États Indépendants et des Balkans, établi par Vincent Hachard, (BIULO) d'après *The World's writing systems* / ed. by Peter D. Daniels and William Bright. New York; Oxford: Oxford University press, 1996. ISBN 0-19-507993-0.
In
MEN-DES-SDBD- Groupe de travail « catalogage des documents en caractères non latins », D. DUCLOS-Faure, janvier 2002,
www.sup.adc.education.fr /bid/Acti/fcnl/fcnl.doc (site consulté le 19 février 2003)

II. Les systèmes de codage informatique des caractères latins étendus et non latins

« En amont de tout système informatisé se situe un élément fondamental dont dépendent toutes les opérations subsidiaires inhérentes aux différentes tâches de traitement y compris les processus de l'indexation, du tri, de la recherche et de la diffusion de l'information. Cet élément de base n'est autre que le principe du codage de l'information » (M. Ben Henda, 38).

Le terme "codage" est défini dans la présentation du standard Unicode comme suit : « Une norme de codage de caractères fournit des unités fondamentales de codage (c'est-à-dire des caractères abstraits) mises en correspondance de un à un avec les numéros de code attribués » (47).

II.1. L'ASCII

L'American Standard Code for Information Interchange (ASCII), créé au cours des années soixante, dans le cadre du développement de l'informatique aux Etats-Unis, a été l'un des tous premiers systèmes de codage de caractères.

Un code de caractère permet d'attribuer un groupe de signaux à un caractère. Il existe des codages internes à l'ordinateur, et des codages internes aux logiciels. Les principes des codages des caractères dont il sera question ne dépendent pas de l'ordinateur.

Toute information est codée à l'aide d'une suite de signaux indiquant le passage ou non de courant. Ces passages et absences de passage de courant étant symbolisés par 1 ou 0, l'information traitée par un système informatique est une suite de 0 et de 1. Un signal est un chiffre binaire ou bit (binary digit) qui a ces deux valeurs possibles. Deux bits représentent donc 2 à la puissance 2, soit 4 valeurs possibles, etc…

L'ASCII étant basé sur 7 bits, le nombre de caractères codés est égal à 2 à la puissance 7, soit 128. Parmi ces 128 caractères figurent les 26 lettres de l'alphabet latin sans accent, en majuscules et en minuscules, des chiffres et des signes divers. Ainsi "A" codé en ASCII est représenté par la suite 1000001, et "a" par 1100001.

Pour coder les 32 caractères cyrilliques utilisés par la langue russe, l'ASCII sur 7 bits est donc insuffisant puisqu'il ne couvre pas, par exemple, les Ч, Ш, Б et Ы.
En fait, la norme ASCII ne satisfait que l'anglais, le shawili et l'indonésien.

II.2. Les autres systèmes de codages

Pour pouvoir coder d'autres écritures, les normes ISO-8859-n (n allant de 1 à 15) ont été élaborées.
Elles sont des extensions de l'ASCII et codent sur 8 bits, c'est-à-dire "2 à la puissance 8", soit 256 caractères. Les 128 premiers sont identiques à ceux de l'ASCII, tandis que les 128 suivants permettent l'écriture d'autres langues.

Il s'agit pour le français des IS0-8859-1 et ISO-8859-15, cette dernière permettant de coder la ligature « e dans l'o ».
L'ISO-8859-5 code les caractères cyrilliques employés par le russe, mais aussi le biélorusse, le bulgare, le macédonien, l'ukrainien et le serbe.
Outre ces normes établies par l'Organisation internationale de Normalisation, d'autres systèmes ont été créés par d'autres instances nationales ou privées[21].

II.3. Les différents codages des caractères cyrilliques

Il existe plusieurs codages des caractères cyrilliques. Certains ont une origine soviétique, d'autres occidentale.

Parmi ces dernières, figurent l'ISO-8859-5 déjà mentionnée et le jeu de caractères CP1251 ou WinCyrillic.

FIGURE N°9 : TABLEAU DES CARACTERES D'APRES LEUR CODE EN NOTATION HEXADECIMALE[22] ISO 8859-5 "CYRILLIC"

code	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
2		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	---
8	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
9	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
A		Ё	Ђ	Ѓ	Є	S	I	Ї	J	Љ	Њ	Ћ	Ќ		Ў	Џ
B	А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П
C	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я
D	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п
E	р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я
F	№	ё	ђ	ѓ	є	s	i	ї	j	љ	њ	ћ	ќ	§	ў	џ

Source : giannieanna.chez.tiscali.fr/codage/tabcar5.htm

[21] Cf Liste non exhaustive de normes de codages informatiques de systèmes d'écritures en annexe 5 p. 93.
[22] La notation hexadécimale consiste à utiliser les chiffres de 0 à 9 et des lettres A à F pour représenter les chiffres de 10 à 15.

FIGURE N°10 : TABLEAU DES CARACTERES D'APRES LEUR CODE EN NOTATION HEXADECIMALE WINDOWS 1251 "WINCYRILLIC"

code	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
2		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	---
8	Ъ	Ѓ	‚	ѓ	„	…	†	‡	€	‰	Љ	‹	Њ	Ќ	Ћ	Џ
9	ђ	‘	’	“	”	•	–	—	□	™	љ	›	њ	ќ	ћ	џ
A		Ў	ў	Ј	¤	Ґ	¦	§	Ё	©	Є	«	¬		®	Ї
B	°	±	І	і	ґ	µ	¶	·	ё	№	є	»	j	S	s	ї
C	А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П
D	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я
E	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п
F	р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я

Source : giannieanna.chez.tiscali.fr/codage/tabcar5.htm

Les codages d'origine soviétique sont marqués par la dénomination GOST, suivie de chiffres, l'acronyme GOST signifiant standard d'État ou encore KOI-8, c'est-à-dire code pour l'échange d'information sur 8 bits. Ils ont évolué dans le temps.
Le jeu de caractères KOI-8 se décline en KOI8-R, pour la langue russe, KOI8-U pour l'ukrainien, le biélorusse…

FIGURE N°11 : TABLEAU DES CARACTERES D'APRES LEUR CODE EN NOTATION HEXADECIMALE KOI8-R

code	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
2		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	---
8	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
9	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
A		╫	╪	ё	╓	╔	╕	╖	╗	╘	╙	╚	╛	╜	╝	╞
B	╟	╠	╡	Ё	╢	╣	╤	╥	╦	╧	╨	╩	╪	╫	╬	©
C	ю	а	б	ц	д	е	ф	г	х	и	й	к	л	м	н	о
D	п	я	р	с	т	у	ж	в	ь	ы	з	ш	э	щ	ч	ъ
E	Ю	А	Б	Ц	Д	Е	Ф	Г	Х	И	Й	К	Л	М	Н	О
F	П	Я	Р	С	Т	У	Ж	В	Ь	Ы	З	Ш	Э	Щ	Ч	Ъ

Source : giannieanna.chez.tiscali.fr/codage/tabcar5.htm

II.4. Incompatibilité des normes de codage

- La norme ISO-8859-5 est incompatible avec la norme KOI,
- La norme KOI8-R et Windows1251 sont incompatibles,
- La norme WinCyrillic est incompatible à la fois avec l'ISO 8859-5 et avec KOI.

D'après Gianna Vacca (15), et comme le montrent les tableaux ci-dessus, à une même lettre correspond un code différent, dans chacun des tableaux. Ainsi la même lettre Ш est codée en BF par la norme KOI8-R, en C8 par la norme ISO 8859-5 et en D8 par le codage Windows 1251.

II.5. Les problèmes posés par l'ASCII et les ISO 8859-n

Les problèmes posés par l'ASCII et les ISO 8859-n sont nombreux et apparaissent à différents niveaux informatiques.

- Dans un logiciel bureautique :

La gestion de deux écritures différentes à l'intérieur d'un même document pose des problèmes dans un environnement bureautique du type Windows 97.
L'explication réside probablement dans le fait que Windows 97 n'ayant pas intégré Unicode, ce système ne dispose que de tables de code pour chaque langue. Cela impose de changer de table et donc de travailler grâce à un mécanisme d'alternance, les langues se substituant les unes aux autres et ne permettant pas la gestion aisée d'un document multiscript (M. Ben Henda, 38). C'est le cas des normes ISO 8859-n qui permettent une gestion alternée mais pas simultanée de différentes langues.

- Dans un message électronique :

L'ASCII présente l'inconvénient de ne pas permettre l'envoi d'un message électronique multiscript si l'une des écritures n'est pas codée par l'ASCII.
Nous avons rappelé en III.1 (Partie I) que l'ASCII ne permet pas de créer des URL en caractères non latins.

- Sur Internet :

L'alternance des écritures codées selon les normes ISO 8859-n entrave la circulation de l'information sur Internet.

Face à cette multitude de systèmes de codage, à ces problèmes de compatibilité, face aux difficultés de gérer un texte comprenant plusieurs écritures et d'échanger des documents, et aussi pour répondre à l'explosion d'Internet, un système de codage universel semble remporter l'unanimité des professionnels. Il s'agit d'Unicode ou ISO-10646.

III. Quelques généralités sur Unicode

N. B : Nous ne ferons qu'effleurer quelques grands principes de la conception et de la construction d'Unicode, essentiellement par le prisme de l'alphabet cyrillique. Nous n'aborderons pas les concepts complexes d'Unicode, comme les notions de caractères combinatoires multiples, qui semblent plutôt liés à des alphabets et à des structures de langues, complexes et très éloignés des langues européennes ou le mécanisme de sérialisation des caractères et le surcodage de transfert qui sont deux des cinq niveaux de représentation des caractères (41).

Unicode est un standard (et non une norme[23]) de codage de caractères. Il concerne plus d'un million de caractères, alphabétiques (dont les caractères cyrilliques), idéographiques, mathématiques... et permet ainsi de coder notamment les écritures européennes, celles du Moyen-Orient qui s'écrivent de droite à gauche et les écritures d'Asie.

III.1. Le consortium Unicode.

Ni la Russie ni aucun autre pays de l'ex-Union soviétique (pas plus que la France, d'ailleurs) ne semble faire partie du Consortium Unicode, alors que l'Inde et le Vietnam, notamment, y sont représentés, par l'intermédiaire de l'International Forum for Information Technology in Tamil et du Technical Committee on Information Technology.

III.2. Les espaces de code de la norme ISO 10646 et du standard Unicode

L'espace de code de la norme ISO 10646 est le domaine des valeurs numériques disponibles pour le codage des caractères sur 32 bits (soit 4 octets).
Pour visualiser cet espace, il faut se représenter un cube dont les arêtes « mesureraient » 2 à la puissance 8. Ce cube peut se voir comme une succession de 256 groupes (1er octet) de 256 plans (2ème octet) de 256 rangées (3ème octet) de 256 cellules (4ème octet).
Chaque plan représente par conséquent un ensemble de caractères codés sur "2 à la puissance 8" fois "2 à la puissance 8", soit "2 à la puissance 16", c'est-à-dire sur 16 bits, soit 65 536 caractères. Le premier plan du premier groupe est appelé le Basic Multilingual Plan (BMP) qui correspond au codage Unicode.

Les normes ASCII, ISO8859-n et le standard Unicode peuvent donc être vus comme des sous-ensembles inclus l'un dans l'autre. Ainsi, les 128 premiers codes du BMP sont ceux de l'ASCII, dont les extensions sont la famille des ISO-8859-n, elles-mêmes incluses dans Unicode.

Il faut préciser que le standard Unicode est identique à la norme ISO 10646 en ce qui concerne le numéro et le nom des caractères, car leurs répertoires et leurs codages sont identiques. Cela résulte de la fusion des deux répertoires en janvier 1992. En fournissant « également d'importants algorithmes de mise en œuvre,

[23] Un standard est un ensemble de règles définies par un ou des groupes privés (exemple : Consortium Unicode) tandis qu'une norme est une règle approuvée par des instances officielles, comme l'ISO (Organisation internationale de normalisation)

des propriétés de caractères et d'autres renseignements sur la sémantique de ceux-ci », Unicode est un outil plus puissant que la norme ISO 10646 (41).

FIGURE N°12 : ZONE DES ECRITURES GENERALES DU BMP (HUIT PREMIERES RANGEES)

00		Latin de base		Supplément Latin 1
01	Latin étendu A		Latin étendu B	
02	Latin étendu B	Alphabet phonétique international	Modificateurs	
03	Signes combinatoires		Grec et copte	
04	**Cyrillique**			
05		Arménien	Hébreu	
06	Arabe			
07	Syriaque		Thâna	

D'après : Introduction à Unicode et à l'ISO 10464 / P. Andriès. (figure 3, p. 58)

Le tableau ci-dessus montre bien que « les espaces de codes sont ordonnés dans une suite qui reproduit, quand c'est possible, la logique évolutive des anciennes normes : le premier espace est réservé à l'ASCII, suivi des caractères latins, grecs, cyrilliques, hébreux, arabes, etc ; conformément à la famille des normes ISO 8859-n » (M. Ben Henda, 38)

Les principes de conception d'Unicode sont au nombre de 10, résumés dans le tableau ci-dessous.

FIGURE N°13 : LES 10 PRINCIPES DE CONCEPTION D'UNICODE

Principe	Énoncé
Codes de caractère à 16 bits	Les codes de caractère Unicode ont une largeur de 16 bits, ils forment un seizet.
Efficacité	L'analyse et le traitement d'un texte Unicode sont simples.
Caractères et non glyphes	Le standard Unicode code des caractères et non des glyphes.
Sémantique	Les caractères ont une sémantique bien définie.
Texte brut	Le standard Unicode code du texte brut.
Ordre logique	La représentation implicite en mémoire est l'ordre logique.
Unification	Le standard Unicode unifie les caractères identiques à l'intérieur des écritures quelles que soient les langues.
Composition dynamique	On compose les formes accentuées dynamiquement.
Séquence équivalente	Chaque forme statique précomposée a une suite équivalente de caractères dynamiquement composés.
Convertibilité	Une convertibilité exacte est garantie entre le standard Unicode et d'autres normes largement répandues.

Source : Unicode 3.1 et ISO 10646 en français
http://iquebec.ifrance.com/ hapax/? (site consulté le 16 juillet 2003)

III.3. Les trois formes de stockage d'Unicode

Le standard Unicode propose trois formes de stockage ou formes en mémoire des caractères (41) à largeur fixe ou variable.

III.3.a/ Les formes à largeur fixe

Ce sont l'UTF[24]-16 et l'UTF-32.

- Forme implicite d'une valeur de 16 bits = UTF-16 (ou UCS-2)

Chaque numéro de caractère est représenté par une suite unique de 16 bits (seizet), considérée et traitée comme une unité de stockage. Elle ne concerne que les caractères du PMB,

- Une forme d'une valeur de 32 bits : UTF-32 (ou UCS-4)

Chaque numéro de caractère est représenté par une quantité sur 32 bits.

III.3.b/ Les formes à largeur variable

Ce sont l'UTF-8 et l'UTF16.

- Une forme d'une valeur de 8 bits = format UTF-8

Il s'agit d'une forme de stockage des caractères, à largeur variable (de un à quatre octets de mémoire) qui permet de répondre aux besoins des systèmes architecturés autour de l'ASCII ou d'autres jeux de caractères à un octet.
Autrement dit, c'est l'une des techniques permettant d'utiliser Unicode par une réduction de longueur de code de manière convenable et compatible avec les environnements développés autour de l'ASCII 7 bits et l'ASCII étendu (codage sur 8 bits). (M. Ben Henda, 38). Elle a été choisie comme forme préférée pour l'internationalisation des protocoles d'Internet (41).

- UTF-16

Chaque numéro est représenté par une suite d'un à deux seizets de mémoire.

	UTF 16	**UTF 32**
Nombre de caractères codés	Plus de 65 000	Plus d'un million

III.4. Distinction caractère / glyphe

Unicode distingue les caractères des glyphes, en ne codant que les premiers.

III.4.a/ Glyphe (ou œil)

Un glyphe est l'image ou encore la trace imprimée d'un caractère qui est une unité minimale ayant un sens (caractère abstrait pour Unicode). Unicode identifie donc le point de code d'un caractère abstrait. Autrement dit, il représente les différentes formes qu'un caractère abstrait peut prendre. Donc un glyphe peut représenter un seul ou plusieurs caractères ou bien plusieurs glyphes peuvent représenter un seul caractère.

Un répertoire de glyphes constitue une police. Ce sont les autres protocoles d'affichage d'un ordinateur qui se chargent du rendu graphique des caractères. En

[24] UTF : Unicode Transformation en anglais ou Format transformé d'Unicode

effet, lors de l'affichage des caractères Unicode, un ou plusieurs glyphes peuvent être sélectionnés pour afficher un caractère particulier. Ces glyphes sont sélectionnés par un moteur de rendu pendant le processus de composition et de disposition.

Exemple :

Glyphe / Oeil	Numéro du caractère	Nom du caractère Unicode / ISO 10646
Ж	U+O416	Lettre majuscule cyrillique JÉ
ж	U+0436	Lettre minuscule cyrillique JÉ
P	U+0420 U+0050	Lettre majuscule cyrillique ERRE Lettre majuscule latine P

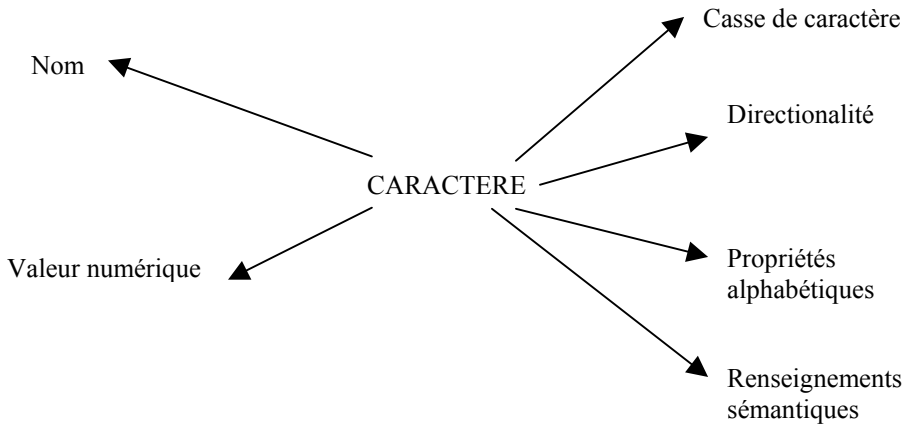
Le glyphe P, connu des francophones, correspond à la lettre majuscule latine P de valeur 0050, dans le tableau des caractères latins de base. Il correspond aussi à la lettre majuscule cyrillique ERRE, dans le tableau des caractères cyrilliques. Donc à un glyphe peuvent correspondre deux lettres de valeurs différentes.

III.4.b/ « Caractère abstrait » et «caractère codé ».

Dans le contexte d'Unicode, il faut distinguer « caractère abstrait » et « caractère codé ».
Un « caractère abstrait » est une unité d'information utilisée pour organiser, commander ou représenter des données textuelles.
Un « caractère codé » est l'association entre un caractère abstrait et son numéro.

III.5. La sémantique des caractères Unicode

Ce standard attribue à chacun des caractères un nom, une valeur numérique et plusieurs autres types de renseignements.



III.5.a/ Nom et valeur numérique

Une valeur de code Unicode se représente à l'aide d'un U+n où n est un nombre composé de quatre à six chiffres en notation hexadécimale.

Exemple : le Ж, dénommé « Lettre majuscule cyrillique JÉ » a la valeur U+0146.

Toute valeur de 16 bits représente toujours le même caractère.

III.5.b/ Autres types de renseignements donnés par Unicode

- La casse des caractères codés (majuscule, minuscule, casse de titre)
- Leur directionalité (gauche, droite, faible, neutre)
- Leurs propriétés alphabétiques
- Des renseignements sémantiques (caractère à symétrie miroir ou pas, numéral ou pas, combinatoire ou pas).

L'ensemble des renseignements fournis par Unicode à chaque caractère représente environ 50 informations par code (C. De Loupy, 18).

Ces renseignements sont rendus nécessaires par la distinction entre « caractère » et « glyphe ». Ils servent à effectuer des tris, des analyses et ils servent aussi à tout traitement par des algorithmes qui ont besoin d'une connaissance sémantique des caractères traités.

III. 6. Notion de texte brut

Un texte brut, pour Unicode, est une suite de codes de caractères. Cette notion s'oppose à celle de texte enrichi ou texte de fantaisie qui est constitué de texte brut et d'informations dont l'identification de la langue du texte, la taille du corps, la couleur du texte, des hyperliens etc..

Unicode ne code que du texte brut, c'est-à-dire qu'il ne définit pas de méthode qui permettrait de distinguer les données codées par Unicode des autres informations supplémentaires.

III.7. Principe d'unification

Le standard Unicode évite de coder plusieurs fois le même caractère. Cela signifie qu'il affecte un seul numéro de code aux caractères utilisés par plusieurs langues au sein d'une même écriture. Autrement dit, chaque lettre commune, signe de ponctuation, symbole et diacritique se voit attribuer un seul code, quelle que soit la langue.

Selon ce principe, « Unicode unifie les caractères identiques à l'intérieur des écritures et non à l'intérieur d'un même alphabet associé à une seule langue » (41).

IV. Les avantages d'Unicode

L'apparition d'Unicode correspond aux besoins :

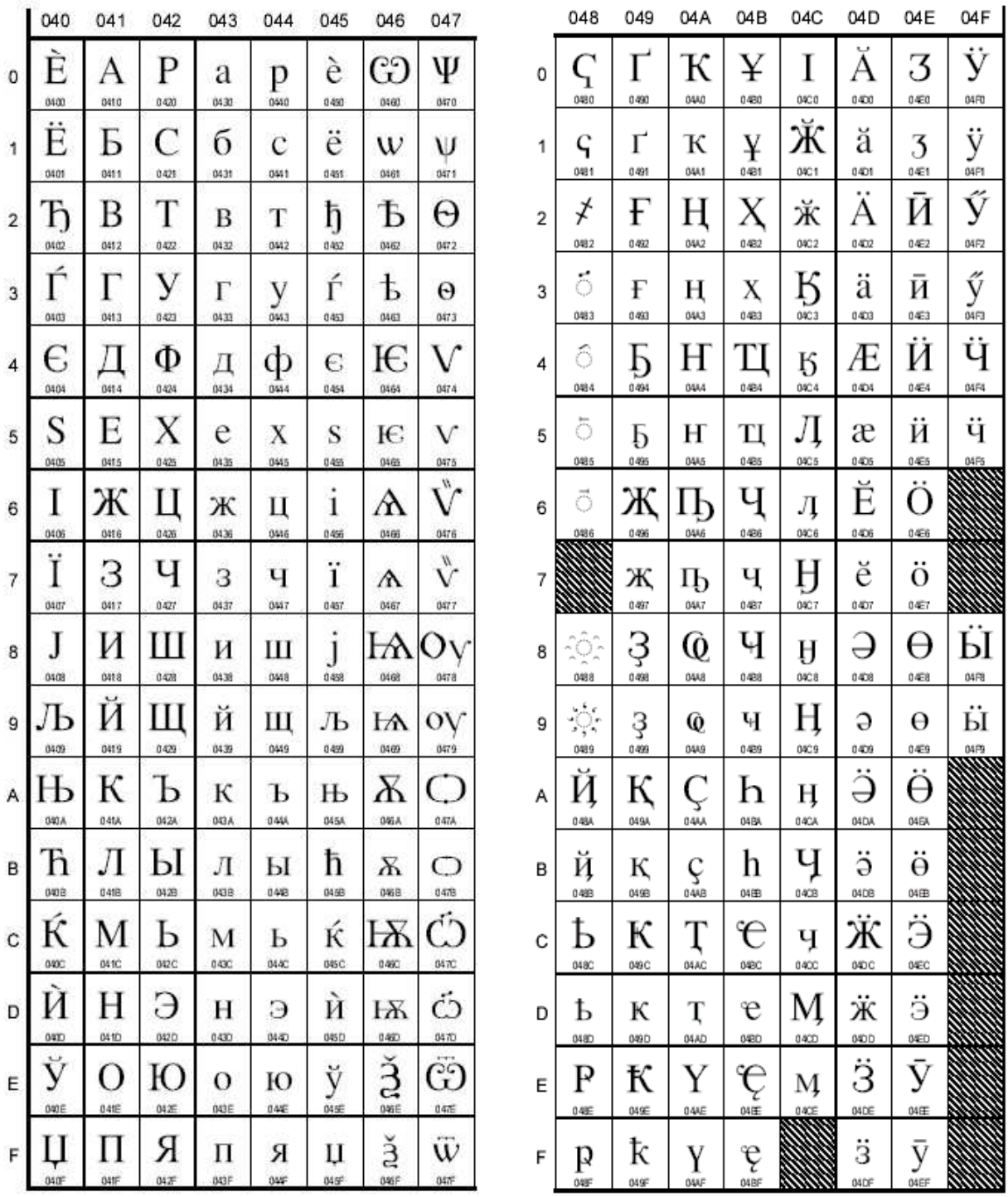
- de lever l'ambiguïté des polices qui utilisent souvent les mêmes valeurs pour coder des caractères et des symboles totalement différents,
- de supprimer l'incompatibilité des différents codes provenant de normes nationales contradictoires,
- de supprimer l'incompatibilité des systèmes d'échanges et d'affichage des données texte (40).
- de rendre transparente la manipulation de toutes les langues à la fois sans ambiguïté de codes et surtout sans obligation de conversion entre tables de codes (38).
- de gérer un texte à plusieurs écritures, ce qui est essentiel, notamment pour la construction d'un système de gestion de bibliothèque dont les fonds sont multilingues et multi-écritures. En codant les caractères d'un très grand nombre d'écritures, Unicode supprime le problème de passage entre deux jeux de caractères dans un même texte multilingue.
- d'indexer et de traiter l'information automatiquement (40).
- de stocker et de manipuler des documents sous leur forme d'origine, c'est-à-dire dans des codages différents. (40).
- d'éviter la prolifération des jeux de caractères et de simplifier le développement de logiciels, en réduisant les coûts, puisque l'adoption d'une norme en favorisant la dissémination du processus industriel, induit une baisse des coûts des composants et des logiciels.
- de disposer d'un jeu de caractères normalisé utilisable à l'échelle planétaire.
- Unicode incorpore des normes bibliographiques comme l'ISO 5426 (47).
- « Unicode fournit un codage qui s'adapte à une grande gamme d'algorithmes de manipulation de texte » (47).
- Unicode fournit également des tableaux de correspondance de casse ou de conversion entre les répertoires de jeux de caractères internationaux ou nationaux (41).

De fait, si la famille des ISO 8859-n continue d'être en vigueur, la substitution par Unicode est en cours et le passage définitif est imminent (38).

V. Le codage du cyrillique par Unicode

Le bloc Unicode de l'écriture cyrillique repose sur l'ISO 8859-5.
Les caractères cyrilliques occupent la rangée 0400-04FF du BMP. Cette rangée peut se présenter sous la forme d'un tableau dont les lignes sont repérées par les chiffres de 0 à 9 puis par les lettres de A à F et les colonnes par des notations hexadécimales : 040 à 049 puis de 04A à 04F, soit 256 points de code, présentés dans le tableau ci-dessous.

FIGURE N°14 : TABLEAU DE CODES UNICODE 3.2 : LETTRES CYRILLIQUES



Source : Unicode 3.1 et ISO 10646 en français
http://iquebec.ifrance.com/ hapax/? (site consulté le 16 juillet 2003)

Alphabet russe de base

0410 А LETTRE MAJUSCULE CYRILLIQUE A
0411 Б LETTRE MAJUSCULE CYRILLIQUE BÉ
→ 0183 Ƃ lettre minuscule latine b potence
0412 В LETTRE MAJUSCULE CYRILLIQUE VÉ
0413 Г LETTRE MAJUSCULE CYRILLIQUE GUÉ
0414 Д LETTRE MAJUSCULE CYRILLIQUE DÉ
0415 Е LETTRE MAJUSCULE CYRILLIQUE IÉ
0416 Ж LETTRE MAJUSCULE CYRILLIQUE JÉ
0417 З LETTRE MAJUSCULE CYRILLIQUE ZÉ
0418 И LETTRE MAJUSCULE CYRILLIQUE I
0419 Й LETTRE MAJUSCULE CYRILLIQUE I BREF
= 0418 И 0306 ◌̆
041A К LETTRE MAJUSCULE CYRILLIQUE KA
041B Л LETTRE MAJUSCULE CYRILLIQUE ELLE
041C М LETTRE MAJUSCULE CYRILLIQUE EMME
041D Н LETTRE MAJUSCULE CYRILLIQUE ENNE
041E О LETTRE MAJUSCULE CYRILLIQUE O
041F П LETTRE MAJUSCULE CYRILLIQUE PÉ
0420 Р LETTRE MAJUSCULE CYRILLIQUE ERRE
0421 С LETTRE MAJUSCULE CYRILLIQUE ESSE
0422 Т LETTRE MAJUSCULE CYRILLIQUE TÉ
0423 У LETTRE MAJUSCULE CYRILLIQUE OU
→ 0478 Ꙋ lettre majuscule cyrillique ouk
→ 04AF ү lettre minuscule cyrillique ou droit
0424 Ф LETTRE MAJUSCULE CYRILLIQUE EFFE
0425 Х LETTRE MAJUSCULE CYRILLIQUE KHA
0426 Ц LETTRE MAJUSCULE CYRILLIQUE TSÉ
0427 Ч LETTRE MAJUSCULE CYRILLIQUE TCHÉ
0428 Ш LETTRE MAJUSCULE CYRILLIQUE CHA
0429 Щ LETTRE MAJUSCULE CYRILLIQUE CHTCHA
042A Ъ LETTRE MAJUSCULE CYRILLIQUE SIGNE DUR
042B Ы LETTRE MAJUSCULE CYRILLIQUE YÉROU
= i dur majuscule
042C Ь LETTRE MAJUSCULE CYRILLIQUE SIGNE MOU
042D Э LETTRE MAJUSCULE CYRILLIQUE É
042E Ю LETTRE MAJUSCULE CYRILLIQUE IOU
042F Я LETTRE MAJUSCULE CYRILLIQUE IA

0430 а LETTRE MINUSCULE CYRILLIQUE A
0431 б LETTRE MINUSCULE CYRILLIQUE BÉ
0432 в LETTRE MINUSCULE CYRILLIQUE VÉ
0433 г LETTRE MINUSCULE CYRILLIQUE GUÉ
0434 д LETTRE MINUSCULE CYRILLIQUE DÉ
0435 е LETTRE MINUSCULE CYRILLIQUE IÉ
0436 ж LETTRE MINUSCULE CYRILLIQUE JÉ
0437 з LETTRE MINUSCULE CYRILLIQUE ZÉ
0438 и LETTRE MINUSCULE CYRILLIQUE I
0439 й LETTRE MINUSCULE CYRILLIQUE I BREF
= 0438 и 0306 ◌̆
043A к LETTRE MINUSCULE CYRILLIQUE KA
043B л LETTRE MINUSCULE CYRILLIQUE ELLE
043C м LETTRE MINUSCULE CYRILLIQUE EMME
043D н LETTRE MINUSCULE CYRILLIQUE ENNE
043E о LETTRE MINUSCULE CYRILLIQUE O
043F п LETTRE MINUSCULE CYRILLIQUE PÉ
0440 р LETTRE MINUSCULE CYRILLIQUE ERRE
0441 с LETTRE MINUSCULE CYRILLIQUE ESSE
0442 т LETTRE MINUSCULE CYRILLIQUE TÉ
0443 у LETTRE MINUSCULE CYRILLIQUE OU
0444 ф LETTRE MINUSCULE CYRILLIQUE EFFE
0445 х LETTRE MINUSCULE CYRILLIQUE KHA
0446 ц LETTRE MINUSCULE CYRILLIQUE TSÉ
0447 ч LETTRE MINUSCULE CYRILLIQUE TCHÉ
0448 ш LETTRE MINUSCULE CYRILLIQUE CHA
0449 щ LETTRE MINUSCULE CYRILLIQUE CHTCHA
044A ъ LETTRE MINUSCULE CYRILLIQUE SIGNE DUR
044B ы LETTRE MINUSCULE CYRILLIQUE YÉROU
= i dur minuscule
044C ь LETTRE MINUSCULE CYRILLIQUE SIGNE MOU
→ 0185 ƅ lettre minuscule latine sixième ton
044D э LETTRE MINUSCULE CYRILLIQUE É
044E ю LETTRE MINUSCULE CYRILLIQUE IOU
044F я LETTRE MINUSCULE CYRILLIQUE IA

Source : Unicode 3.1 et ISO 10646 en français
http://iquebec.ifrance.com/ hapax/? (site consulté le 16 juillet 2003)

Il est à noter que les caractères utilisés par l'alphabet russe de base (majuscules et minuscules comprises) n'occupent que les quatre colonnes 041, 042, 043 et 044. Les autres colonnes comprennent, notamment, des lettres historiques et du cyrillique étendu.

Les caractères latins faisant partie de l'alphabet cyrillique ne sont pas recodés en cyrillique (47). Ce qui illustre l'un des principes de conception du standard Unicode qui est l'unification.

Cinq caractères de l'alphabet russe de base sont signalés par des descriptions particulières, dans la liste de noms de caractères qui suit le tableau de code. Il s'agit de :

0411 Б LETTRE MAJUSCULE CYRILLIQUE BÉ
→ 0183 ƃ lettre minuscule latine b potence

Cette description indique un renvoi vers la lettre latine b potence. Il s'agit d'indiquer que les deux caractères ne sont pas identiques bien que leurs glyphes sont fort similaires.

0419 Й LETTRE MAJUSCULE CYRILLIQUE I BREF
= 0418 И 0306 ◌̆

Cette description indique une « décomposition canonique », ce qui signifie que le caractère Й se décompose en un И et un signe diacritique.

0423 У LETTRE MAJUSCULE CYRILLIQUE OU
→ 0478 Ꙋ lettre majuscule cyrillique ouk
→ 04AF ү lettre minuscule cyrillique ou droit

Deux renvois sont signalés pour le caractère У dont un vers la lettre majuscule « ouk » qui classée par Unicode dans les « lettres historiques ».
L'autre renvoi concerne un caractère cyrillique étendu utilisé en azéri et en bashkir, notamment.

042B Ы LETTRE MAJUSCULE CYRILLIQUE YÉROU
= i dur majuscule

L'indication fournie ici est le nom le plus usité du caractère.

044C ь LETTRE MINUSCULE CYRILLIQUE SIGNE MOU
→ 0185 ƅ lettre minuscule latine sixième ton

Ce renvoi vers une lettre latine indique également que les deux caractères ne sont pas identiques bien que leurs glyphes sont fort similaires.

VI. Les limites d'Unicode

Unicode ne codant que du texte brut, il ne précise ni la taille, ni le format, ni encore l'orientation des caractères à l'écran, il ne tente pas de coder des caractéristiques de texte comme la langue, la police, la force de corps, l'emplacement etc…

Par conséquent, certains aspects typographiques sont omis dans ce standard et diverses manières d'écrire un caractère cyrillique dans des langues données peuvent ne pas apparaître (13).

Bien que deux auteurs annoncent que peu de langages de programmation supportent Unicode en natif (40), (ce qui signifie que peu de langages de programmation ont été créés avec ce standard), le site d'Unicode fait savoir, lui, que ce système de codage « est exigé par de nombreux standards récents tels que XML, Java, JavaScript, LDAP, CORBA 3.0, WML, etc.

L'un des inconvénients que présente l'utilisation d'Unicode est le doublement possible de la taille des fichiers, ce qui risque de poser des problèmes de circulation de ces fichiers sur les réseaux (39) et de remettre en cause l'ensemble des mécanismes informatiques en place. C'est plus précisément le codage avec UCS-4 qui est gourmand en place, car il stocke tous les caractères sur 4 octets, même si les deux premiers octets sont systématiquement à zéro. La forme Utf-8, elle, quant à elle, « un mécanisme de stockage relativement compact en termes d'octets » (41).

Selon H. Hudrisier (43), « tant que cet environnement à 8 bits subsistera, les nouveaux usages auront du mal à émerger, parce que même si des mécanismes d'adaptation ont été prévus (UTF-8), un environnement aussi ancien et mal dimensionné provoquera immanquablement d'innombrables dysfonctionnements». Ce qui amène cet auteur à envisager un développement graduel de l'utilisation d'Unicode : « la mise en place d'une informatique et de réseaux à 16 bits suppose des aménagements fondamentaux par rapport aux systèmes à 8 bits (réécriture de logiciels systèmes, des protocoles de télécommunication, évolution technologique des machines) qui plaide pour que la norme qui succède à l'ASCII soit directement prévue sur 32 bits (4 octets soit 4 milliards de codes théoriques), même si sa mise en œuvre informatique doit être progressive».

Nous avons déjà pu constater combien Unicode est un apport précieux dans le développement du multilinguisme sur la Toile. Son rôle de socle pour tout développement multilingue dans un système informatique et dans le monde des bibliothèques se confirme à la lumière des recommandations faites par les experts.

Notons enfin que les problématiques liées aux caractères indiens ou arabes semblent inspirer les auteurs d'articles traitant du multilinguisme et d'Unicode, plus que ne le font les caractères cyrilliques. Parmi les documents et ouvrages traitant de ce sujet, peu de références en russe[25] ont été trouvés.

[25] International Forum On Information And Documentation. Vol. 23 N°4. 1998
017044 - INFOethics 98 : the multilingual environment in Cyberspace / Grytsenko, Volodymyr; Anisimov, Anatoly. - (p. 6-10) ; Kasparova, N.N. (1998), "Cataloguing of documents for multilingual catalogues fo libraries in Russia : analysis of the problem situation", in Byrum, J.D. Jr and Madisaon, O (Eds), *Multi-script, Multi-lingual, Multi-character Issues for the Online Environment*, K.G. Saur, Munich.

Troisième partie : Propositions et conclusions

I. Propositions pour l'évolution de la construction de l'annuaire des chercheurs de défense

La construction de cet annuaire pourrait évoluer selon plusieurs directions, de façon concomitante. Ces directions concernent la structure de la base elle-même, le système de codage des caractères, et les questions de romanisation.

I. 1. Les champs

Concernant les champs de la base , il paraît pertinent de :

- supprimer le champ « date de naissance » de l'expert qu'il conviendrait d'indiquer, le cas échéant, dans le champ "Renseignements complémentaires",

- indiquer une variante de la transcription des noms propres dans le champ "Nom" et non pas dans celui des « Renseignements complémentaires ». Cela permet, lors de l'interrogation par l'utilisateur, d'augmenter les chances de trouver la fiche (ce qui n'est pas le cas si la transcription alternative est placée dans le champ "Renseignements complémentaires" puisque celui-ci n'est lié à aucun autre champ de la table).

I.2. Les tables :

Afin de permettre une recherche plus simple dans la base et de découpler la recherche de chercheurs de celle d'un centre, il serait utile de créer une table intitulée "Annuaire des centres de réflexion" avec des champs supplémentaires : "Statut du centre et son financement", "langue(s) du site", "Sigle du centre". Dans ce champ, serait indiqué le sigle figurant en caractères latins dans l'URL du site.

Dans le même esprit, c'est-à-dire, afin de découpler la recherche de publications de celle de chercheurs ou de centres, il serait intéressant de créer une table intitulée "Annuaires des périodiques" avec les champs : "Directeur de la publication", "type d'abonnements", "langues de la publication", et « URL de la publication en ligne ».

Ces deux tables supplémentaires seraient liées entre elles et liées à « l'Annuaire des chercheurs de défense ».

I.3. Le codage des caractères

Afin d'assurer le codage de tous les caractères des langues des chercheurs de l'annuaire, il serait utile de faire migrer cette base sous Access 2000 qui a intégré Unicode ou du moins d'installer Unicode dans la version utilisée. Le but est de pouvoir réintroduire des caractères originaux pour les noms des chercheurs, le nom des centres de réflexion et les titres des publications et de disposer ainsi d'une base bilingue français-russe et à terme multilingue. Cependant, il conviendrait de s'assurer qu'il est possible, tant d'un point de vue technique qu'économique, de saisir aisément tous les caractères des écritures employées dans ces pays.

Nous devons, à ce propos, signaler que les deux stagiaires arabisant et sinisant ayant effectué le même type travail afin d'alimenter l'annuaire, à partir des ressources Web n'ont été confrontés ni au problème des caractères originaux, ni à celui de la romanisation, leurs sources étant souvent anglophones.

I.4. La romanisation

Il conviendrait également, dans l'optique d'une base multilingue, de conserver un système de romanisation, qui pourrait être la translittération ISO 9/95. Il serait donc nécessaire de sensibiliser les personnes qui seront chargées de saisir ces données en caractères latins étendus et de mettre en place un référentiel interne de translittération, afin d'assurer une homogénéité la plus grande possible au sein de l'annuaire. Ce souci ne correspond pas seulement à une vision idéaliste de documentaliste pointilleux(se) mais à la nécessité de fournir à l'utilisateur final un ensemble de résultats le plus cohérent possible.

II. Présentation de recommandations et de solutions plus générales

Des chercheurs et spécialistes de la documentation et des questions de multilinguisme proposent des solutions et des recommandations afin de faciliter l'élaboration de documents multi-écritures et multilingues, d'améliorer leurs échanges via un système informatique. Ces solutions sont liées à Unicode et concernent les divers formats de catalogage et la romanisation des noms d'auteurs et de collectivités.

II.1. Windows et Unicode

Unicode est mis en œuvre dans tous les systèmes d'exploitation et langages informatiques, comme Java, Javascript (ECMAScript), MS Windows 2000 et XP, Mac OS/X (41).
Ces versions sont donc à privilégier par rapport aux anciennes qui n'avaient pas encore intégré Unicode. Mais il est toutefois possible d'y implanter Unicode en effectuant des opérations informatiques indiquées en annexe 1 p.78.

II.2. Codage d'un document numérique en format HTML

Pour coder un document HTML qui comprend des caractères non latins, la règle essentielle consiste à signaler au lecteur le codage utilisé. Il faut pour cela ajouter une ligne après la balise <HEAD> qui est :

<META HTTP-EQUIV="Content-Type" CONTENT="text/html ;charset= >

et indiquer après « charset » le codage utilisé.

Il faut aussi :

- un jeu de caractères approprié à la langue du document et le plus standard possible ;

- taper autant que possible tous les caractères dans ce jeu de caractères, sans utiliser de références d'entité (exemple &eacute qui code le e accent aigu) ou numériques ;
- transmettre le document avec une étiquette de codage de caractères (Charset MIME) appropriée. (34)

II. 3. Le catalogage des ressources électroniques : Dublin Core

Dans le domaine du catalogage des ressources électroniques, un projet allant vers une unification pour créer un système de référencement universel a vu le jour au milieu des années quatre-vingt-dix, il s'agit de Dublin Core. Selon M. Ben Henda (38), cette unification « constitue l'une des alternatives les plus importantes à la description unifiée des ressources d'informations électroniques, en présentant « un début de consensus international afin d'harmoniser cet effort de référencement universel, consolidé par l'avènement Unicode ». Il est adopté par de grandes structures spécialisées dans toutes les disciplines.
Pour S. Haigh (8) « il semble bien peu probable qu'il remplace MARC en ce qui a trait aux descriptions de ressources bibliographiques où l'utilisation de MARC est bien implantée et durable ».

Dublin Core comporte quinze éléments de données sur les données (métadonnées) qui peuvent être contenues dans la ressource électronique qu'il décrit. Lorsque les métadonnées sont incorporées à la ressource, elles sont nommées selon deux conventions qui sont la syntaxe d'étiquette MÉTA HTML et le cadre de définition des ressources RDF (Resource Description Framework). RDF étant une expression du langage XML (8), le lien entre Dublin Core et XML est manifeste.

II.4. XML et Unicode

Aujourd'hui, les formats MARC semblent susciter moins d'engouement auprès des spécialistes du monde documentaire qui s'intéressent davantage à d'autres formats comme XML (Extended Markup Language).

Les arguments plaidant en faveur de ce format sont nombreux. Tout d'abord, il est basé sur Unicode, ce qui permet donc d'éliminer tout problème de codage de caractères.
Le langage XML est une réplique du langage HTML en plus perfectionné, dans le sens où il définit un algorithme sans aucune ambiguïté pour déterminer le jeu de caractères utilisé par le codage. Dans le langage XML, un attribut de codage facultatif sur la déclaration XML définit le codage des caractères. L'algorithme suivant détermine les codages par défaut : si le fichier commence par une marque d'ordre d'octet Unicode [0xFF 0xFE] ou [0xFE 0xFF], le document est considéré comme étant en codage UTF-16. Dans le cas contraire, il est en UTF-8 (39).

Par ailleurs, il présente un intérêt notable pour le monde des bibliothèques puisqu'il permet de fournir une définition précise de la structure d'un type de documents numériques et offre des possibilités d'arborescence illimitée. Cette préférence pour XML s'exprime à la BnF par son adoption pour l'élaboration de

plusieurs produits comme Gallica et les Signets de la BnF, car « XML permet de traiter de façon similaire tout ce qui peut être publié, qu'il s'agisse de données bibliographiques ou de documents numériques » (A. Garden, 6).
XML est donc un des formats préconisés notamment, pour le monde de la documentation. Il a comme avantage de pouvoir décrire des documents multimédias, volumineux et à structure complexe.

II. 5. Romanisation et Unicode

L'opération de romanisation a pour but de permettre au catalogueur et à l'usager, ne connaissant pas la langue originale du document, d'élaborer, de trouver et de lire la notice de ce document. Tous deux ont besoin de connaître le système de romanisation choisi : « Quand une bibliothèque romanise ou cyrillise, le chercheur doit le savoir, doit connaître le schéma de conversion employé et doit être capable de l'utiliser pour définir sa clé de recherche. Un chercheur peut ne pas être au courant des pratiques de la bibliothèque et utiliser un schéma tout à fait différent. » (7). Avec le développement des catalogues en ligne, l'incertitude dans laquelle peut se trouver un chercheur risque d'être d'autant plus fréquente que le chercheur sera amené à faire sa requête, de plus en plus souvent, non pas sur le lieu même de la bibliothèque mais en interrogeant le catalogue en ligne de la bibliothèque.

II. 5.a/ Les recommandations du CLENOL

Les recommandations du groupe des experts cyrilliques, piloté par D. Duclos-Faure (11), et travaillant dans l'optique de la création de la BULAC[26], recouvrent les trois volets : affichage, interrogation, et saisie.

- Les notices bibliographiques doivent être affichées en caractères originaux.
- L'interrogation doit pouvoir se faire et en caractères originaux et en translittération.
- La saisie se fait en translittération. Les systèmes sont capables de retranscrire en caractères originaux pour les faire apparaître à l'affichage. (le produit Loris de la société Ever-Team, par exemple, intègre une procédure de conversion d'une écriture dans l'autre pour ne faire qu'une saisie). Outre le fait qu'une seule saisie est nécessaire, la retranscription en caractères originaux permet à un locuteur de la langue de contrôler la bonne saisie de la translittération.

En matière de normes, les recommandations sont les suivantes (11) :

- Application de la norme ISO 9 enrichie 1995 pour le cyrillique,
- Mais aussi privilégier les caractères originaux cyrilliques,
- Double enregistrement des zones en écriture originale et en romanisation (et identification précise du système de romanisation).

Un des intérêts que présente la norme ISO 9/95 réside dans le fait de pouvoir « passer automatiquement d'un système d'écriture à l'autre. Elle s'impose donc

[26] BULAC : Bibliothèque Universitaire de Langues et de Civilisations.
Anciennement : Pôle Langues et civilisations du monde jusqu'en 2002.

d'elle-même dans un système automatisé où l'alimentation des index nécessite des zones en caractères latins » (5).

En ce qui concerne la romanisation et la translittération, il convient d'affiner ces considérations un peu générales, pour se pencher sur les cas des formes courantes. Noms, prénoms de personnes et nom de collectivités ne sont pas traités sur le même plan.

II. 5.b/ Le cas des prénoms

Un prénom ne peut pas être traité comme un nom de famille, l'usage étant prédominant sur l'automaticité d'une norme. De plus, il est fortement déconseiller de créer une forme courante qui n'est pas attestée par l'usage.
L'exemple des deux prénoms russes Сергей et Юрий en est une bonne illustration. Le prénom de l'un des fondateurs de Google, Сергей, est transcrit sur ce site sous la forme Sergey. On en déduit que Y est la transcription du Й. Si l'on respecte cette règle automatiquement, le prénom Юрий, lui, peut être transcrit sous la forme Yuriy, qui n'est pas attestée, la forme courante adoptée par les anglo-saxons étant Yuri.
Le Groupe d'experts pour le cyrillique du CLENOL (5) et F. Hours-Richard, notamment, recommandent l'élaboration d'une liste de prénoms russes sous leur forme française attestée par l'usage. L'établissement de cette liste permettrait, notamment, d'éviter les formes anglicisées.

II. 5.c/ Le cas des collectivités

En ce qui concerne les noms de collectivités, toujours d'après le groupe d'experts pour le cyrillique, il n'y a pas de transcription courante. Seules les formes traduites en français peuvent être considérées comme des formes courantes. Le groupe recommande de prendre « le nom d'une collectivité dans la langue utilisée couramment par cette collectivité sur ses publications ».

II. 5.d/ L'adoption des caractères originaux

L'adoption des caractères originaux dans les notices est une préconisation souvent faite par différents auteurs. Parmi les raisons avancées figurent les suivantes :

- L'alphabet latin est d'un usage exclusif dans moins de la moitié du monde (9),

- Chacun doit pouvoir accéder à un catalogue dans la forme qui lui convient et donc à l'aide d'écritures non latines. En effet, le résultat attendu par l'usager est l'accès à l'information recherchée, quelle que soit la forme utilisée dans la requête, romanisée ou non.

En matière d'adoption des caractères originaux, on constate une contradiction entre les préconisations de règles en vigueur et celles du Groupe de travail du ministère de l'Éducation Nationale. Alors que « les règles internationales de description bibliographiques (ISBD) préconisent le plus souvent l'utilisation par une agence bibliographique donnée, d'une seule langue de catalogage et / ou une seule écriture » (13), le groupe de travail "Catalogage des documents en caractères non latins préconise, lui, pour le Pôle langues et civilisations du monde,

la réalisation d'un catalogue unique contenant les caractères originaux non latins et des caractères latins, c'est-à-dire un catalogue multi-écriture (11).

II. 5.e / Vers la disparition de la romanisation ?

Si les trois points préconisés par le CLENOL semblent faire l'unanimité parmi les professionnels du catalogage, l'éventuelle disparition de la romanisation que l'introduction d'Unicode pourrait rendre caduque, ne remporte pas tous les suffrages. En effet, certains souhaiteraient voir abandonnée ou du moins sévèrement limitée la translittération pour les points d'accès, dans la mesure où Unicode permettra la gestion de plusieurs écritures dans un même système (10). Mais, H. Hudrisier (43), quant à lui, estime que « Nous ne devons en aucun cas considérer que la translittération est une étape dépassée. »

Si le procédé de translittération n'est donc pas à abandonner, la coexistence de multiples formes de romanisation devrait connaître une évolution vers trois grandes types d'écritures utilisées dans les notices bibliographiques, à savoir :

- La transcription de la Library of Congress,
- La translittération selon les normes ISO,
- Les caractères originaux.

Il n'est pas exclu que certaines institutions procèdent à un panachage de deux solutions, comme la BnF qui tend à utiliser à la fois les caractères originaux et la translittération ISO. Cette démarche s'inscrit bien dans la pensée de H. Hudrisier (43) pour qui « il est donc indispensable et réaliste de continuer à produire des banques de données bibliographiques par translittération des titres et des noms d'auteurs mais il serait aussi beaucoup plus commode de pouvoir les inscrire dans leurs langues respectives ».

Nous n'avons pas jusqu'à présent abordé le problème que peut poser la saisie erronée des données par le catalogueur. La probabilité de commettre des erreurs lors de la saisie de données sous une forme romanisée augmente lorsque le catalogueur ne connaît pas la langue d'origine. La solution technique émanant d'un ancien catalogueur et présentée ci-dessous a l'avantage de palier des manques de connaissances linguistiques des catalogueurs et de rallier tous les avis : « Une solution générale au problème des inexactitudes de transcription dans les notices bibliographiques consiste à utiliser des liens hypertexte. Dans un catalogue pour le Web, il est possible d'avoir un lien à une image numérisée. Une image numérisée ne permet toutefois pas de rechercher une occurrence spécifique d'un glyphe donné, mais il s'agit là d'une opération qui concerne plutôt les recherches en texte intégral que sur le catalogue » (7).

III. Les réponses des éditeurs de SIGB et propositions pour le cahier des charges du Centre documentaire du Campus de Défense.

Dans le cadre de son développement, le Centre Documentaire du Campus de Défense a besoin d'acquérir un système intégré de gestion de bibliothèque (SIGB), capable de fonctionner avec un certain nombre de caractères non latins, dont les cyrilliques. En amont de la phase d'élaboration du cahier des charges, une grille d'analyse a été soumise à des éditeurs de SIGB.

Celle-ci a été constituée à partir de la grille extraite du rapport du groupe de travail « Catalogage des documents en caractères non latins et latins étendus » (11) que nous avons augmentée. Ce travail avait été effectué avant de découvrir les divers problèmes présentés ci-dessus. C'est pourquoi nous nous attacherons, dans ce qui va suivre, à suggérer au Centre documentaire du Campus de Défense, des questions complémentaires à rajouter à une nouvelle grille ou dans le cahier des charges qui sera adressé aux éditeurs de SIGB.
Parmi les sept éditeurs, contactés en février et en septembre, quatre nous ont renvoyé une grille remplie. Et tout d'abord, il convient de présenter brièvement et par ordre alphabétique les deux sociétés qui ont souhaité être citées (les deux autres sont appelées « éditeur XX » et « éditeur YY »).

III. 1. Présentation de sociétés éditrices de SIGB

III.1.a/ Ever Team

Ever Team est une société française qui emploie plus de deux cents personnes. Ses bureaux sont situés à Paris, Grenoble et Lyon. Ses filiales sont basées en Espagne, en Allemagne, au Canada et aux Etats-Unis.
Parmi ses clients, figurent des partenaires de langues non latines (Moyen-Orient, Maghreb). Les produits d'Ever Team occupent une part de marché importante dans les administrations françaises dont la Direction Générale pour l'Armement (ministère de la Défense), les collectivités locales et le domaine de l'Enseignement. La part de marché d'Ever Team s'élève à 60 % dans les universités.
Un des produits phares de la société est le logiciel Loris (gestion de bibliothèque) dont le club utilisateurs compte une centaine de membres.

III.1.b/ Ex Libris

Ex Libris est une société internationale qui édite le logiciel Aleph 500. Celui-ci est en service dans plus de 700 bibliothèques ou centres de documentation dans le monde, dont la BDIC, le CERN[27] qui avait adopté le système Aleph dès 1992 et qui devait s'équiper du logiciel Aleph 500 en 2001.

III.2. Étude des réponses des éditeurs de SIGB

Les réponses ont été rassemblées dans un même document, afin de faciliter leur comparaison (Cf. annexe 4, p. 82).

N.B 1: la société XX n'ayant pu communiquer qu'une partie des réponses à la grille d'analyse pour finir à temps l'étude comparative des réponses, nous avons néanmoins travaillé avec celles qu'elle nous avait communiquées.

N.B 2 : Les points relevés ici ont fait l'objet de réponses différentes qui deviennent ainsi elles-mêmes sources d'interrogation. Les autres, ayant rassemblé des réponses homogènes, sont supposés ne pas poser de problème majeur.

[27] Conseil Européen de la Recherche Nucléaire

III. 2.a/ Les insuffisances du clavier français et les claviers multilingues

Nous avons déjà constaté que, d'une part, la saisie des lettres Č, Š, Ž, et Ŝ est nécessaire pour effectuer une translittération rigoureuse, selon la norme ISO 9/95 et que d'autre part le clavier français ne s'y prête pas aisément, puisque les touches correspondantes n'existent pas. Pour remédier à cette absence de touches, il existe des solutions plus ou moins ergonomiques dont la création de touches de raccourci (Cf. annexe 3, p. 81).
Cependant, une fois ces touches de raccourci créées, on constate que cette solution n'est pas valable pour insérer ces lettres dans une table du logiciel Access et pas davantage dans l'interface Web d'un moteur de recherche.

Les éditeurs proposent, pour certaines écritures, des solutions de copier-coller ou de claviers virtuels.

Il existe aussi d'autres solutions comme celles proposées par Mysoft[28]. Le logiciel TwinLink fournit une police spécifique et un clavier-écran pour 60 langues.
La page "L'écriture cyrillique" du site www.culture.fr/culture/edm/fr/Cy/cy03.html indique d'autres outils de saisie comme Tango Creator, Uniedit, Unitype et TwinBridge.
Le clavier virtuel multilingue pourrait peut-être permettre de saisir des signes diacritiques dans une interface Web. Nous ne l'avons pas testé, faute de logiciel approprié.

Les inconvénients posés par un clavier ne possédant que les touches correspondantes à l'alphabet latin ne se limitent pas au monde la documentation et des bibliothèques. Des scientifiques peuvent aussi être confrontés à des problèmes de clavier, dans certains contextes particuliers, comme celui décrit par un article du quotidien Le Monde [29]: « Malgré les moyens fantastiques du nouveau supercalculateur de NEC, il y a encore des choses à améliorer et "*il faut relativiser les performances actuelles*", précise Gervan Madec. Pour commencer, il n'est pas évident pour un scientifique non japonais de travailler sur un clavier japonais… »

III.2.b/ Les moteurs de recherche

Les réponses relatives aux moteurs de recherche intégrés dans les SIGB des éditeurs XX et YY ne permettent pas d'affirmer que tous les jeux de caractères sont pris en compte par ces moteurs. L'éditeur XX indique que les langues dans lesquelles se fait l'interrogation en texte intégral sont celles gérés par le système, c'est-à-dire, les langues d'écritures latines, grecque et cyrilliques. Dans le cas de l'éditeur YY, il conviendrait de savoir si le moteur RetrievalWare gère d'autres langues que les latines, annoncées (à savoir : le français, l'anglais, l'allemand, l'espagnol et l'italien).

[28] Mysoft : www.mysoft.fr
[29] Earth Simulator, un monstre de plus de cinq mille processeurs pour mimer les phénomènes qui animent et agitent la Terre / C. Galus.-
In : Le Monde (Paris, 1944) = ISSN 0395-2037.- (2003-02-26) n° 18234

III.2.c/ L'affichage

En ce qui concerne l'affichage, l'éditeur YY est le seul à annoncer que son produit ne permet pas l'affichage en caractères romanisés, diacrités et originaux.

III.2.d/ La gestion des prêts

La gestion des prêts, via la messagerie électronique est réalisée soit à l'aide de la messagerie déjà en place dans la structure, soit à l'aide d'un système intégré au SIGB. Pour l'instant seuls Ever Team et Ex Libris semblent en mesure de réaliser une messagerie multilingue.

III.2.e/ Unicode, Unimarc et les caractères non latins

En matière d'indexation, de stockage des données, de création de bases de données, d'échanges de fichiers, d'interface Web, le standard Unicode est manifestement la solution choisie par les éditeurs. Néanmoins, il apparaît que d'une part, tous les SIGB ne gèrent pas tous les caractères non latins et que d'autre part, il existerait une incompatibilité entre Unimarc et Unicode, alors que ces deux systèmes, l'un codant des caractères, l'autre définissant des formats de données bibliographiques, n'ont, a priori, rien en commun.. Cela se manifeste dans les points suivants :

- L'éditeur XX n'est pas en mesure de manipuler les caractères arabes et asiatiques, tandis que l'éditeur YY ne manipule que ceux prévus par Unimarc dont Iso 54 26 qui ne correspond qu'au jeu de caractères latins étendus.
- L'éditeur YY évoque, par exemple, la possibilité d'utiliser des « identifiants Unicode » si le caractère n'est pas dans le jeu Unimarc.
- Les éditeurs affirment que le format de stockage des données bibliographiques qu'ils utilisent est compatible avec Unicode.
- Quant à la prise en compte des caractères non latins par Unimarc, les réponses sont peu convaincantes. Celle de l'éditeur YY incite à consulter la nouvelle édition d'Unimarc.
- L'importation de fichiers en langues étrangères et dans des formats différentes fait entrer en jeu la norme ISO 2709, citée par ExLibris et l'éditeur YY. Ever Team, lui, évoque d'autres jeux : ANSEL, ISO 5426 et Unicode.

Par ailleurs, Joan M. Aliprand dans son article « Le catalogage dans un univers à plusieurs écritures : les limites » (7) renforce cette impression, en suggérant qu'Unimarc définit ses propres répertoires de caractères.

Ce point mériterait d'être approfondi et éclairci. Cela aurait demandé davantage de recherches documentaires et de temps. Les réponses obtenues auprès d'experts ne nous ont, malheureusement, pas permis de lever toutes les ambiguïtés. Nous tentons néanmoins d'aborder cette interrogation et d'en dresser le cadre.

- Un rapide aperçu de l'histoire des formats MARC

Le projet d'élaborer un catalogue lisible en machine fut confié à Henriette D. Avram, en 1964, par la Bibliothèque du Congrès de Washington. Le format MARC est une traduction informatique des normes bibliographiques. Les premières

notices MARC ne concernent que des monographies, en anglais, et imprimées aux Etats-Unis. Puis la British National Bibliography compose son propre format (BNB-Marc) complètement compatible avec l'américain. Ensuite divers formats MARC se mettent à proliférer. L'ISO qui met alors en œuvre la norme ISO 2709, au milieu des années soixante-dix, dénonce le danger de cette prolifération, dans la perspective de partages d'informations entre différents systèmes. C'est aussi alors qu'apparaît le futur Unimarc, format international dont une des caractéristiques « est la possibilité de faire des liens entre les notices » (14).
Parmi les formats MARC nationaux qui fleurissent à partir du milieu des années soixante-dix, citons le français Intermarc et le russe Rusmarc. Dix ans après sa première publication, Intermarc est entièrement révisé pour prendre en compte l'évolution d'Unimarc. Sa spécificité réside dans le fait qu'Intermarc concerne uniquement la Bibliothèque Nationale (14).

Certains auteurs, comme Yves Desrichard ou Pierre-Yves Duchemin, cité par Annie Garden (6), semblent estimer, à la fin des années quatre-vingt-dix, que le format MARC n'est plus adapté aux exigences d'un contexte multimédia en évolution rapide, et que le passage à d'autres formats, comme SGML, est nécessaire : « pour ces derniers [formats MARC], l'avènement de bases de données relationnelles ou des formats de type SGML, où la notion de lien est largement plus développée et en quelque sorte « native » à leur élaboration, montre bien que les limites d'un type de format vieilli, peu adapté aux évolutions informatiques les plus récentes » (Y. Desrichard, 10). Annie Garden (6), quant à elle, évoque les limites des formats MARC « qui invitent la profession à rejoindre le monde du World Wide Web et à utiliser d'autres formats ».

D.R. Millier, cité par D. Lahary (2) évoque, lui, la complexité inutile du format MARC. D. Lahary en fait l'illustration en remarquant que « la codification des caractères, sujet douloureux s'il en est, a fait l'objet de solutions spécifiques aux formats MARC, et parfois à un seul d'entre eux : pour une notice bibliographique en alphabet latin fournie en UNIMARC, il faut se référer à la norme ISO 5426, solution qui diffère de celle de MARC21 ».

- UNIMARC et les notices multiscript

La question du multiscript dans une notice au format UNIMARC met en évidence la complexité évoquée plus haut.
En 1999, J.M. Aliprand (7) mettait en garde contre un trop grand enthousiasme pour Unicode et notamment pour son utilisation simultanée avec Unimarc. Y. Desrichard (42) renchérit peu de temps après, en écrivant que « les formats MARC ne sont pas réellement adaptés à la gestion de plusieurs écritures au sein d'une même notice ». Le problème d'inadaptation du format Marc à la gestion de plusieurs écritures au sein d'une même notice bibliographique réside dans le codage : « Unicode fournit un répertoire d'écritures et de signes bien plus grand que ceux prévus par le systèmes de bibliothèques que ce soit ceux qu'utilisent USMarc[30] ou Unimarc » (J.M. Aliprand, 7). Dans la version 3.0 du standard Unicode, 237 caractères cyrilliques sont codés, tandis que UNIMARC n'a à sa disposition que 102 caractères. Ce qui signifie que des caractères codés par Unicode pourraient ne pas figurer dans une notice Unimarc.

[30] USMarc ou Marc21

Nombre de signes prévus par MARC et Unicode

Écriture	USMARC / Unimarc	Unicode version 3.0
Cyrillique	102	237
Latin	21	163
Arabe	124	141
Idéogrammes de l'Asie de l'Est	13 469	27 484

Source : Le catalogage dans un univers à plusieurs écritures : les limites / Joan M. Aliprand.- 65th IFLA Council and General Conference, Bangkok, Thailand, august 20 – august 28, 1999 www.ifla.org/IV/ifla65/papers/079-155f.htm (site consulté le 25 juin 2003)

De plus, les répertoires de caractères d'Unimarc définissent parfois des encodages propres pour des éléments qui sont unifiés dans UNICODE (7). D'où il pourrait résulter des problèmes de cohabitation entre les deux systèmes.

Toujours d'après Joan M. Aliprand (7), le répertoire de caractères Unicode est néanmoins officiel pour l'UNIMARC. Mais, les difficultés pourraient surgir lors d'échanges : « S'il est tout à fait possible aujourd'hui de créer à la source des bases de données bibliographiques dans les formats MARC sous Unicode, il est toutefois difficile d'en assurer l'échange car les données Unicode dans les enregistrements devraient être au préalable converties dans un codage 8 bits de Unimarc ou USMarc. Il n'est pas encore évident de savoir comment introduire Unicode comme répertoire unique de caractères pour tous les formats Marc du monde» (M. Ben Henda, 38).

Le bon fonctionnement de la norme Unimarc et du standard Unicode et leur coexistence au sein d'un même système seront, par conséquent, des points à mettre en exergue lors de la rédaction du cahier des charges, en vue de l'acquisition d'un SIGB.
D'après les réponses de ces éditeurs, Ever Team et ExLibris semblent maîtriser le mieux les questions de gestion des caractères non latins dans un SIGB.
Cependant, des aspects non ou peu abordés peuvent faire l'objet de questions complémentaires à soumettre aux éditeurs.

III.3. Quelques suggestions pour compléter la grille d'analyse

III.3.a/ La norme ISO 9/95, le catalogage et l'indexation

Il est très peu question de normes ISO de translittération dans la grille d'analyse de SIGB. Il serait nécessaire de s'assurer que le SIGB pourra intégrer ces normes sans difficulté, car l'intérêt de ces normes ISO réside dans le fait de pouvoir « passer automatiquement d'un système d'écriture à l'autre. Elle s'impose donc d'elle-même dans un système automatisé où l'alimentation des index nécessite des zones en caractères latins » (5). Un index en caractères latins permet, quant à lui, de n'effectuer qu'une requête dans une seule langue, même si le terme recherché est présent dans la base de données en plusieurs langues.

Il conviendrait par conséquent de rajouter à la grille d'analyse les questions suivantes :

1. Le SIGB peut-il intégrer les normes ISO de translittération pour les langues employées par le Centre documentaire du Campus de Défense ?
2. Un document pourra-t-il être rédigé à l'usage des catalogueurs, afin de leur indiquer les normes et les systèmes de romanisation à utiliser, et les séquences de saisie des lettres spéciales et des diacritiques ?
3. Peut-il y avoir des problèmes de saisie de signes diacritiques, lors du catalogage ?

III.3. b/ Saisie des caractères diacrités par l'usager

Lors d'une requête effectuée par l'usager, sur place ou via une interface Web, faudra-t-il interdire à l'interrogation, l'emploi des signes diacritiques ? [Si oui, ne pas oublier de le signaler à l'écran sur la page, à côté du masque de saisie de la requête].

III.3.c/ Affichage des caractères diacrités par l'interface Web

Existe-t-il un procédé technique propre au SIGB permettant d'afficher correctement les signes diacritiques via l'interface Web, quelle que soit la plate-forme de l'usager (Macintosh, PC) ?

III.3.d/ Les claviers virtuels

L'acquisition de claviers virtuels est-elle possible pour toutes les langues des documents détenus par le Centre documentaire du Campus de Défense ?
Le passage d'un clavier à l'autre sur le même poste informatique est-il aisé ?

III.3.e/ La norme MIME et les échanges de messages électroniques multiscripts

La norme MIME est-elle intégrée au système de messagerie du SIGB ?

III.3.f/ Version d'Unicode

Quelle version d'Unicode est prévue dans le SIGB ? La dernière en date ? Comment pourra se faire la mise à jour lors de la sortie de la version suivante ?
N.B. : La version 4. 0 d'Unicode devrait sortir à l'automne 2003 (41).

III.3.g/ Unimarc et Unicode

Quelle est la dernière version d'Unimarc prévue pour le SIGB ? Quels jeux de caractères y sont prévus ? Sont-ils compatibles avec Unicode ?

Le Centre documentaire du Campus de Défense devant rentrer dans le réseau du SUDOC, nous signalons que le passage du SUDOC à Unicode et en relation, extension de l'utilisation d'Unimarc, est prévu pour 2004-2005.

IV. Projets des bibliothèques russes et ex-soviétiques et échanges internationaux

Parmi la littérature recueillie sur les activités des bibliothèques et les projet d'informatisation menés en ex Union soviétique, signalons trois projets ayant fait l'objet d'une coopération internationale.
Il s'agit de l'informatisation de la Bibliothèque d'État de Russie, menée dans le cadre d'un programme TACIS[31], de SONEGOS, portail des bibliothèques

[31] TACIS : Technical Assistance to the Commonwealth of Independent States.

nationales de la CEI et le projet d'informatisation des archives du Komintern, sous l'égide du Conseil de l'Europe.

IV. 1. L'informatisation de la Bibliothèque d'État de Russie.

L'intitulé exact du projet pilote, financé par TACIS est la « Création d'un système d'informatisation pour la Bibliothèque d'État de Russie ». Le budget était d'un million d'euros pour une durée de dix mois.

Cette institution comprend 42 millions d'ouvrages et des publications en 249 langues vivantes et mortes. La première tentative d'informatisation de ce qui s'appelait alors Bibliothèque d'État V.I.Lénine, date de la fin des années soixante. Elle échoue en raison de la complexité de la tâche et du manque d'expérience dans ce domaine. Une second échec intervient en 1990, mais cette fois, davantage, pour des raisons financières. Les actions menées à partir des années 1996 aboutissent notamment à la création d'un réseau de fibres optiques entre bibliothèques, et qui a rendu possible l'accès à Internet.

Le projet TACIS, lancé à partir de fin 1997, comprenait notamment (1) :

- la formation de quinze personnes, chargées de former elles-mêmes les autres membres du personnel de la bibliothèque.
- l'élaboration d'un cahier des charges pour un appel d'offres de progiciel de système intégré de bibliothèque et son acquisition. Le logiciel choisi, Aleph 500 de l'éditeur Ex Libris, gère les caractères cyrilliques. Toute la documentation et les écrans (interfaces, aides) relatifs à ce logiciel ont été traduits en russe.
- la conversion des données du système MEKA vers un catalogue en ligne OPAC. Ces données sont constituées des acquisitions russes depuis mai 1998, de livres étrangers acquis depuis 1999 et une base de données de thèses et de résumés de thèses. Cet OPAC contenait 500 000 notices en 2000.
- une rétroconversion en format USMARC d'un échantillon de 10 000 fiches du catalogue d'ouvrages en russe du 19ème siècle par une société mixte russo-allemande (ProSoft-M).

Les réunions entre experts de l'Union européenne et ceux de la Bibliothèque d'État ont été animées lorsqu'il s'est agi du futur format de catalogage de la bibliothèque. Le format retenu a été USMARC, car le format MARC y était déjà utilisé[32].

[32] Le document contenant ces informations est une fort mauvaise traduction qui ne m'a pas permis de comprendre toutes les raisons de ce choix.

IV. 2. Le site SONEGOS

Le site SONEGOS[33], bilingue Russe / Anglais, est un portail de sites de bibliothèques de la CEI ((www.rsl.ru/SONEGOS). Il a été créé en 2000, avec pour modèle le site GABRIEL, qui est son équivalent européen. Les informations concernant les fonds, les collections, les catalogues et les services des bibliothèques sont remises à jour régulièrement.
Le processus de collecte des informations concernant toutes ces bibliothèques a été long et complexe, selon M. Shvartsman, chef du département Réseau de la Bibliothèque d'État de Russie. Cette collecte s'est effectuée notamment grâce au lancement d'un concours du meilleur site de bibliothèque et au dépouillement des questionnaires envoyés aux bibliothèques à cette occasion.
La politique linguistique suivie par les bibliothèques des différents pays de la CEI suit celle de leurs gouvernements. Trois grandes lignes se dégagent à l'examen des sites par M. Shvartsman. La première consiste à n'utiliser que la langue nationale (Ukraine, Moldavie), la seconde adopte systématiquement l'anglais et le russe (Kazakhstan, Belarus), la troisième opte pour la langue nationale et l'anglais (Arménie et Géorgie). L'association des bibliothèques d'Ouzbékistan quant à elle, rajoute le russe à la langue nationale et à l'anglais, tandis que la bibliothèque régionale de l'est du Kazakhstan propose quatre langues : le russe, l'anglais, l'allemand et le kazakh (4).

IV. 3. Le projet Incomka

Ce projet, dont un des partenaires est la Direction des Archives de France, a consisté à numériser et à informatiser une partie des archives du Komintern (ou 3ème Internationale).
Les autre partenaires sont regroupés au sein du Comité international d'informatisation des archives du Komintern qui comprend notamment des représentants du Service des Archives d'État de la Fédération de Russie (ROSARKHIV), des archives d'État pour l'Histoire sociale et politique de la Russie (RGASPI), des Archives fédérales de l'Allemagne et de la Bibliothèque du Congrès de Washington (28).
Lancé en 1996, Incomka devait être achevé dans le courant du premier semestre 2003. L'inauguration de la salle de lecture du RGASPI à Moscou ainsi que l'ouverture à la recherche de la base de données et des images des documents dans les sites des partenaires devaient avoir lieu en juillet dernier (30).
Les produits du projet, à savoir « une base de données regroupant les renseignements contenus dans l'ensemble des inventaires du fonds du Komintern (22 000 pages) et un million de documents numérisés, choisis par un comité scientifique » (29) devront être remis à la Direction des Archives de France.
Mais la numérisation d'un million de pages ne représenterait que 5% du fonds d'archives du Komintern. Le logiciel permettant l'inventaire complet informatisé est Archidoc.

[33] SONEGOS : il s'agit des deux premières lettres des trois mots en russe composant l'abréviation SNG (Sodružestvo Nezavisimyh Gosudarstv)

Les langues qui apparaissent dans ce fonds sont majoritairement le russe et l'allemand, mais une cinquantaine d'autres y seraient également employées.
Afin que les chercheurs non russophones puissent interroger la base de données, il a fallu procéder à une translittération des caractères cyrilliques en caractères latins des noms de famille. Ce qui a été réalisé par la Bibliothèque du Congrès à Washington (29).

Conclusion

Les bibliothèques et les centres de documentation, à la fois confrontés et partie prenante dans les échanges et la coopération au-delà des barrières linguistiques, jouent un rôle essentiel dans la création, la gestion et la diffusion de l'information. Aussi, ils pourraient influer, en faisant valoir leurs exigences auprès des constructeurs de matériels et des éditeurs sur les derniers points d'inégalité restants, en matière de traitement linguistique, à savoir la saisie, les polices, et les traitements de texte. Leur rôle dans un traitement d'égalité quasi totale du plus grand nombre possible d'écritures quant au codage de leurs caractères pourrait ainsi prendre de l'ampleur et concourir à l'augmentation de la variété culturelle, linguistique et de la production des idées.

L'idée présentée par la Directrice des services techniques de l'American university in Cairo Library (9) en est une illustration. Il s'agit de la création d'une centrale de notices d'autorité auteurs multilingues et multi-écritures, destinée à favoriser l'échange international de sources d'information dont chacun pourrait tirer les éléments utiles à son propre catalogue, avec une recomposition possible de la notice d'autorité en fonction des besoins du contexte local, tout en maintenant l'obligation de se conformer aux normes internationales en vigueur.

L'un des enjeux de cette mise à égalité complète du plus grand nombre possible d'écritures dans le monde informatique est la mise en valeur, via Internet, par exemple, des autres cultures et langues que l'anglais.

D'après C. Fluhr[34], l'information multilingue a pris beaucoup d'importance aux Etats-Unis depuis le milieu des années quatre-vingt-dix. Au moins 50% de l'information présente sur la Toile n'est pas rédigée en anglais. Les explications sont diverses et relèvent de plusieurs domaines différents. L'apparition du commerce électronique aurait favorisé l'émergence de l'idée selon laquelle vendre dans la langue du client peut être une approche commerciale efficace. La veille dans le domaine de la recherche scientifique peut être également plus productive si elle est orientée vers des langues autres que l'anglais car il arrive que la recherche ne soit pas diffusée en anglais. Par ailleurs, cette langue ne serait pas suffisante non plus dans les échanges au sein des groupes multinationaux. Enfin, les événements du 11 septembre 2001 auraient renforcé les activités de renseignement militaire aux Etats-Unis et celles-ci nécessiteraient d'accéder à des sources dont la langue n'est pas l'anglais.

Mais une constatation s'impose : l'informatique multiscripturale n'en est qu'à ses débuts. Elle n'est pas encore totalement transparente, ni entièrement automatisée puisque des solutions sont encore à développer comme la traduction automatique et que des périphériques comme les claviers posent encore des problèmes.

[34] Conférence lors du salon iexpo, le 18 juin 2003

Á ces questions techniques, s'ajoutent les aspects humains, c'est-à-dire les aspects culturels de l'échange et du traitement d'un très grand nombre de langues. Selon H. Hudrisier (43), le Web et Unicode ne sont en effet pas suffisants pour assurer la mise en place d'un système dans lequel s'échangerait et serait traité un très grand nombre de langues. Il va devenir nécessaire d'« inventer un tourisme linguistique », de « vulgariser les écritures du monde », et de développer un « métier de l'échange multiscriptural » afin de préserver la variété des identités linguistiques et culturelles.

Dans cette optique, porter un regard très attentif aux activités des autres agences bibliographiques, pas seulement américaines, faire un travail de veille sur le lancement de projets internationaux impliquant des pays de langues non latines, notamment, permettrait d'intensifier les échanges de l'Union européenne avec ces pays et de favoriser ainsi le développement d'outils multi-écritures et multilingues. Mais ici, la dynamique à mettre en œuvre relève, en premier lieu, davantage des domaines politique et culturel que technique.

BIBLIOGRAPHIE ANALYTIQUE

Cette bibliographie est organisée en fonction des grands thèmes traités dans le mémoire, eux-mêmes classés par ordre alphabétique.

Bibliothèques

N°	Références bibliographiques
1	Créer un système d'information pour la Bibliothèque d'Etat de Russie : un projet pilote défiant les technologies de l'information / Monica Segbert, Alexander Vislyi. – 66th IFLA Council and General Conference, Jerusalem, Israel, 13-18 august. www.ifla.org/IV/ifla66/papers/056-142f.htm (site consulté le 4 juillet 2003)
2	Informatique et bibliothèques : vers la banalisation des outils ? / D. Lahary.- In : BBF 2002, Paris, t.47, n°1. p. 60-66.
3	Multilinguisme et multiscripts, l'avenir informatique / Marie-Renée Cazabon. – [2] p. In : BBF 2002, tome 47, N°6, p.106-107
4	The SONEGOS Website as a Gateway to the Libraries of the Commonwealth of Independent States / M. Shvarstman. – [5] p. In : IFLA Journal = ISSN 0340-0352. - (2002) vol. 28: n°2, p.69-73

Catalogage

N°	Références bibliographiques
5	Catalogage des caractères non-latins et latins étendus / Groupe d'experts pour le cyrillique. Synthèse des travaux. Janvier 2002 Document interne.
6	L'avenir des formats de données / Annie Garden.- In : BBF 2001 – Paris, t.46, n°5, p.114-116. http://bbf.enssib.fr/html/2001_46_5/2001-5-p114-garden.xl.asp?print=1 (site consulté le 1 septembre 2003)
7	Le catalogage dans un univers à plusieurs écritures : les limites / Joan M. Aliprand.- 65th IFLA Council and General Conference, Bangkok, Thailand, august 20 – august 28, 1999 www.ifla.org/IV/ifla65/papers/079-155f.htm (site consulté le 25 juin 2003)

8	Le projet de métadonnées Dublin Core / Susan Haigh. In : Flash réseau n°63 ISSN 1200-5304 Services de technologie de l'information, Bibliothèque nationale du Canada http://www.nlc-bnc.ca/9/1/pl-262-f.html (site consulté le 7 juillet 2003)
9	Les autorités auteurs arabes dans l'environnement informatisé aujourd'hui : options et implications / M. Spiers Plettner. www.uni.bamberg.de/unibib/melcom/Plettner-franz.html (Site consulté le 24 juillet 2003)
10	Les formats et normes de catalogage / Yves Desrichard. – [13] p. – Bibliog. In : BBF 1998 –Paris, t.43,n°3, p.56-65) http://bbf.enssib.fr/html/1998_43_3/1998-3-p56-desrichard.xml.asp?print=1 (site consulté le 1 septembre 2003)
11	MEN-DES-SDBD- Groupe de travail « catalogage des documents en caractères non latins », D. DUCLOS-Faure, janvier 2002, www.sup.adc.education.fr /bid/Acti/fcnl/fcnl.doc (site consulté le 19 février 2003)
12	RCAA 2R Interprétation de règles, 22.3C2 – personnes mises en vedettes à leur nom de famille, www.nlc-bnc.ca/6/18/s18-213-f.html (Site consulté le 28 juillet 2003)
13	UNIMARC : quelles solutions pour le catalogage en plusieurs langues et plusieurs écritures ? / Freyre, Elisabeth ; Bourdon, Françoise. – 65th IFLA Council and General Conference. Bangkok, Thailand, 20-28 août 1999 www.ifla.org/IV/ifla65/papers/100-155f.htm (site consulté le 30 juin 2003)
14	Unimarc, Manuel de catalogage / M-R Cazabon.- Ed. du Cercle de la Librairie, 1993

Codage et jeux de caractères

N°	Références bibliographiques
15	Codage informatique des systèmes d'écriture Version 2.0.1 / Gianna Vacca http://giannieanna.chez.tiscali.fr/codagecharsets201.pdf (site consulté en février 2003)
16	Le traitement informatique des documents en caractères non latins / A. Dupas. In : BBF, Paris,T.43, n°1, 1998 http://bbf.enssib.fr/bbf/html/1998_43_1/1998-1-p104-dupas.xml.asp (site consulté en février 2003)

17	Le traitement informatique des documents en caractères non latins : la solution envisagée par le SCD Lyon 3 et d'autres exemples / Amélie Dupas. – 1996-1997. – Rapport de stage : Diplôme d'études supérieures spécialisées Traducteur-Documentaliste Scientifique : Université de Pau et des pays de l'Adour : 1996-1997. www-scd.univ-lyon3.fr/doclec/dupas.htm (site consulté en février 2003)
18	Multilinguisme et document numérique : la dimension technique à l'épreuve du codage des caractères / Claude de Loupy. – http://biblio-fr.info.unicaen.fr/bnum/jelec/Solaris/d06/6loupy.html (site consulté le 29 mai 2003)
19	Prise en charge du caractère unicode www.microsoft.com/windows2000/f…/sag_DNS_und_UnicodeCharacterSupport.htm (site consulté le 1 septembre 2003)
20	Qu'est-ce que MIME ? / Guy Brand www.chimie.u-strasbg.fr/membres/GB/MIME.html (site consulté le 31 mai 2003)
21	The Cyrillic Charset Group / Roman Czyborra http://czyborra.com/charsets/cyrillic.html#GOST-19768-87 (site consulté le 16 juillet 2003)
22	What is KOI-R ? http://koi8.pp.ru/framed-koi8.html (site consulté le 9 juillet 2003)

Écritures

N°	Références bibliographiques
23	Écritures du monde www.culture.fr/edm/fr/index.html (site consulté le 2 juillet 2003)
24	L'aventure des écritures http://classes.bnf.fr/dossiercr/je-code.htm (site consulté le 31 mai 2003)
25	L'écriture cyrillique www.culture.gouv.fr/culture/edm/fr/Cy/cy03.html (site consulté le 2 juillet 2003)

Editeur MtScript

N°	Références bibliographiques
26	L'édition des textes multilingues / Abdel-Malek Boualem, Stéphane Harié. http://rist.cerist.dz/ArticlesFullText/8-1/htm/BOUALEM.htm (site consulté le 3 juillet 2003)

Informatique générale

N°	Références bibliographiques
27	Raccorder son réseau d'entreprise à Internet / A. Fenyö, F. Le Guern, S. Tardieu. – Eyrolles, 1999 . 513 p.

Projet Incomka (Les archives du Komintern)

N°	Références bibliographiques
28	Incomka Comité international pour le projet d'informatisation des archives du Komintern www.ica.org/old/fr/mb/projets.html#incomka (site consulté le 7 juin 2003)
29	Incomka l'Internationale numérisée / H. Ochanine. - Archimag N°161, février 2003
30	Programme « Informatisation des archives du Komintern / Conseil de l'Europe www.coe.int/T/F/Coop%E9ration_culturelle/Culture/Assistance_&_Développement/Archives/activi t%E9s.asp#P49_1995 (site consulté le 7 juin 2003)

Romanisation, transcription, translittération

N°	Références bibliographiques
31	Guide des difficultés de rédaction en musique / Marc-André Roberge www.mus.ulaval.ca/roberge/gdrm/01-trans.htm (site consulté le 18 août 2003)
32	Information et documentation. Translittération des caractères cyrilliques en caractères latins. Langues slaves et non slaves. – Edition juin 1995. - Paris : AFNOR, 1995. – 14 p. – ISSN 0335-3931

Toile et le multilinguisme

N°	Références bibliographiques
33	Approches structurelles et multilingues de l'accès matière sur le Web / CHAN, Lois Mai; XIA, Lin; MARCIA, Lei Zeng. – 65th IFLA Council and General Conference. Bangkok, Thailand, 20-28 août 1999 www.ifla.org/IV/ifla65/papers/012-117f.htm (site consulté le 30 juin 2003)
34	HTML en toutes langues http://babel.alis.com:8080/web_ml/html/index.html (site consulté le 16 juillet 2003)
35	HTTP international http://babel.alis.com:8080/web_ml/html/index.html (site consulté le 16 juillet 2003)
36	Internationalisation du langage de balisage hypertexte (HTML) http://babel.alis.com:8080/web_ml/html/rfc-il _8n-0.html (site consulté le 16 juillet 2003)
37	Projets sur Internet et activités bibliographiques en Russie / Elena Zhabko. – 68th IFLA Council and General Conference. August 18-24, 2002. www.ifla.org/IV/ifla68/papers/065-152f.pdf (site consulté le 4 juillet 2003)

Unicode

N°	Références bibliographiques
38	Apport Unicode, html4 et Dublin Core pour le traitement du texte multilingue : le cas des langues arabo-latines / Mokhtar BEN HENDA.- IX Sommet de la francophonie (9, 2001, Beyrouth, Liban). www.chez.com/benhenda/publicat/indexation_arabe.htm (site consulté le 16 juin 2003)
39	Documentation technique sur l'Unicode / Vincent Colnat. – Maison de l'Orient et de la Méditerranée –Jean Pouilloux CNRS –Université Lyon 2.- Décembre 2001. www.mom.fr/bdd/grecancien/docUNICODE.pdf (site consulté le 1 septembre 2003)

40	Intégration d'Unicode Conception d'un agent de recherche d'information sur Internet / E. Giguet, N. Lucas. In : *Unicode, écriture du monde ?* [Texte imprimé] / sous la dir. de Jacques André, Henri Hudrisier .Dir. 2002 Paris : Hermès. Dans: Document numérique (Texte imprimé) ; vol. 6, nº3-4, 2002. 364 p. : ill., couv. ill. en coul. ; 24 cm Numéro de : "Document numérique", ISSN 1279-5127, vol. 6, nº3-4, 2002. - Bibliogr. en fin de chapitres ISBN: 2-7462-0594-7
41	Introduction à Unicode et à l'ISO 10464 / P. Andriès. In : *Unicode, écriture du monde ?* [Texte imprimé] / sous la dir. de Jacques André, Henri Hudrisier .Dir. 2002 Paris : Hermès. Dans: Document numérique (Texte imprimé) ; vol. 6, nº3-4, 2002. 364 p. : ill., couv. ill. en coul. ; 24 cm Numéro de : "Document numérique", ISSN 1279-5127, vol. 6, nº3-4, 2002. - Bibliogr. en fin de chapitres ISBN: 2-7462-0594-7
42	Jeux de caractères : Unicode a le beau rôle / Desrichard Yves. – [3]p. In : Archimag. N°155, juin 2002, p.28-30
43	L'appropriation des technologies multiscripturales / H. Hudrisier.- In : *Unicode, écriture du monde ?* [Texte imprimé] / sous la dir. de Jacques André, Henri Hudrisier .Dir. 2002 Paris : Hermès. Dans: Document numérique (Texte imprimé) ; vol. 6, nº3-4, 2002. 364 p. : ill., couv. ill. en coul. ; 24 cm Numéro de : "Document numérique", ISSN 1279-5127, vol. 6, nº3-4, 2002. - Bibliogr. en fin de chapitres ISBN: 2-7462-0594-7
44	Prise en charge du caractère unicode www.microsoft.com/windows2000/f…/sag_DNS_und_UnicodeCharacterSupport.htm, (site consulté le 1 septembre 2003)
45	The Unicode standard : a global solution to localization problems in electronic documents. / CHAHUNEAU, Francois. - Document numérique. Vol 1 N° 4. Décembre 1997 (p. 385-401)(1 réf.)
46	The Unicode Standard : its scope, design principles, and prospects for international cataloguing / ALIPRAND, Joan M. - Library Resources And Technical Services. Vol. 44 N° 3. July 2000 (p. 160-167)(19 réf.)
47	Unicode 3.1 et ISO 10646 en français http://iquebec.ifrance.com/ hapax/? (site consulté le 16 juillet 2003)

TABLE DES ILLUSTRATIONS ET HORS-TEXTES

GLOSSAIRE

TERME	CONTEXTE	DÉFINITION
AACR (Anglo-American Cataloguing Rules)	Catalogage	Format de catalogage.
ALA-LC	Translittération	Système de translittération américain
Autorité	Catalogage	Forme officielle d'un nom ou d'un énoncé de sujet utilisé comme clé d'accès dans un catalogue.
Bicamérale	Ecriture, alphabets européens	Une écriture bicamérale distingue les lettres majuscules des minuscules.
Caractère	Unicode	(1) Unité de base utilisée par le codage de caractères d'Unicode (2) Synonyme de caractère abstrait
Caractère abstrait	Unicode	Unité d'information utilisée pour organiser, commander ou représenter des données textuelles. Il n'a pas de forme concrète.
Caractère codé	Unicode	Association entre un caractère abstrait et son numéro
Casse	Unicode	(1) Trait de certains alphabets où les lettres ont deux formes distinctes. (2) propriété normative de caractères : majuscule, minuscule, casse de titre.
Casse de titre		lettre initiale majuscule, les autres lettres du mot étant en minuscules
Diacritique		(1) Signe graphique adjoint à un symbole afin de créer un nouveau symbole qui représente une valeur nouvelle ou modifiée (Unicode) (2) Signes graphiques destinés à empêcher la confusion entre des mots homographes
Directionalité	Unicode	Direction de l'écriture : direction ou orientation des caractères écrits au sein de lignes textuelles d'un système d'écriture.
Élement META	Programmation	Balise utilisée par les moteurs de recherche afin de récupérer les informations définissant le contenu des sites.
Espace de code	Unicode	Le domaine des valeurs numériques disponibles pour le codage des caractères
Étiquette	Unicode	Balise linguistique
Fichier d'autorité	Catalogage (BnF)	Ensemble structuré des notices d'autorité par type de notices
GEAC		Système informatique ; adopté par la Bibliothèque Nationale en 1985 puis abandonné
Glyphe	Unicode	Représente les différentes formes qu'un caractère abstrait peut prendre.
HTML	Internet	Hyper Text Markup Language. Langage qui définit le contenu et permet l'affichage des pages html par les navigateurs. Il est possible de visualiser le contenu d'une page html en regardant le code source.

TERME	CONTEXTE	DÉFINITION
HTTP	Internet	Hyper Text Transport Protocol C'est un protocole de transfert de données utilisé par le Web. Un serveur http est chargé d'envoyer les pages html à un ordinateur
IETF	Internet	Internet Engineering Task Force. Organisme qui crée et maintient à jour les standards Internet.
KOI	Codage cyrillique	Kod Obmena Informatsiej Code pour l'échange d'informations
MARC	Catalogage	Machine Readable Cataloguing = Catalogue lisible en machine
MIME Multi-Purpose Internet Mail Extension	Codage informatique	Format standard qui permet d'imbriquer ou de joindre des fichiers de type arbitraire aux messages de courrier électronique ou d'étiqueter les fichiers HTML renvoyés au navigateur.
OCLC Online Computer Library Center	Bibliothèque	Réseau de bibliothèques Anciennement : Ohio College Library Center
Octet	Informatique	8 bits
Ordre logique	Unicode	Ordre dans lequel le texte est saisi au clavier
Rendu	Unicode	1. Processus lié à la sélection et à la disposition de glyphes afin de représenter des données textuelles. 2. Fait de rendre visible des glyphes sur une unité de visualisation.
Répertoire	Unicode	Ensemble de caractères abstraits destinés à être codés
Rétroconversion	bibliothèque	Transformation des catalogues papiers en dossiers numériques et intégration dans un seul dossier qui représente un catalogue.
RFC (Request For Comments)	Internet html, http,	ils décrivent les standards pratiqués sur Internet (protocoles, extensions,…)
RLIN (Research Library Information Network)	Bibliothèque	Réseau bibliographique américain.
Seizet	Unicode	Suite de 16 éléments binaires considérée et traitée comme une unité.
Sibil	bibliothèques	C'est le premier réseau de catalogage partagé de bibliothèques. La Suisse en était à l'origine.
Technologie RFID	Gestion de bibliothèque SIGB	puce « sans contact », dans le cadre d'une bibliothèque, elle permet d'éviter le « douchage » des documents à emprunter. Identification radio fréquence. Étiquette comportant un circuit intégré avec mémoire de 96 bits et communique par radio fréquence.
Texte brut	Unicode	Est constitué d'une suite de codes de caractères, dont aucun ne représente du balisage. Est donc une suite simple de codes de caractères Unicode
Texte enrichi ou texte de fantaisie	Unicode	Un texte enrichi est constitué d'un texte brut et d'un ensemble d'informations (couleur de texte, hyperliens….)
Think tank	Pensée stratégique Défense	Centre de réflexion stratégique
UCS	Unicode	Jeu de caractères universels

TERME	CONTEXTE	DÉFINITION
URI	Internet	Universal Resource Identifier – Extension Internet de la notion de chemin permettant d'accéder à des documents en spécifiant l'adresse réseau complète, identification et authentification comprise, leur chemin, leur nom, et des paramètres d'accès. Comparer avec "chemin".
URL	Internet	Uniformed Ressources Locator. Nom donné à l'adresse Internet d'une page.
UTF	Unicode	Unicode Transformation Formats

Sources :

- Ged à (presque) tout faire / M. Remise. - Archimag, octobre 2002, n°158, p.24-25
- Glossaire de la BnF : site http://www.bnf.fr/pages/zNavigat/frame/infopro.htm?ancre=outibib/fa-notaut.htm, consulté le 5 octobre 2003
- Glossaire Osinet URL : http://www.osinet.fr/code/glo.asp?Initial=L site consulté en juillet 2003
- Glossaire Unicode URL : http://www.unicode.org/glossary site consulté en juillet 2003
- Introduction à Unicode et à l'ISO 10646 / P.Andries.
  In
  Unicode : écriture du monde ? / J. André, H. Hudrisier.- Paris ; Lavoisier, 2002.- Hermès.
- MEN-DES-SDBD- Groupe de travail « catalogage des documents en caractères non latins », D. DUCLOS-Faure, janvier 2002,
  www.sup.adc.education.fr /bid/Acti/fcnl/fcnl.doc (site consulté le 19 février 2003)

Liste des sigles et acronymes

SIGLE ou ACRONYME	DÉVELOPPÉ
AACR	Anglo-American Cataloguing Rules
ABES	Agence Bibliographique de l'Enseignement Supérieur
ASCII	American Standard Code Information Interchange
BMP	Basic Multilingual Plane
BULAC	Bibliothèque Universitaire de Langues et de Civilisations
CLENOL	Groupe de travail sur le catalogage des caractères non latins et latins, mis en oeuvre par la sous-direction des bibliothèques et de la documentation du ministère de l'Éducation nationale.
HTML	Hyper Text Markup Language.
HTTP	Hyper Text Transport Protocol
IETF	Internet Engineering Task Force. Organisme qui crée et maintient à jour les standards Internet.
IFLA	International Federation of Library Associations and Institutions
KOI	(Kod Obmena Informatsiej) Code d'échange d'informations
MARC	Machine Readable Cataloguing
MIME	Multi-Purpose Internet Mail Extension
OCLC	Anciennement : Ohio College Library Center, puis : Online Computer Library Center
PMB	Plan Multilingue de base
RFC	Request For Comments
RLIN	Research Library Information Network
SIGB	Système Intégré de Gestion de Bibliothèque
SUDOC	Système universitaire de documentation
UCS	Jeu de caractères universels
URL	Uniformed Ressources Locator.
UTF	Unicode Transformation Formats

Liste de sites

Agence Bibliographique de l'Enseignement Supérieur (ABES)	www.abes.fr
Bibliothèque nationale de France (BnF)	www.bnf/fr
Bulletin des Bibliothèques de France	http://bbf.enssib.fr
Direction des Archives de France	www.archivesdefrance.culture.gouv.fr
Fondation Nationale des Sciences Politiques	www.sciences-po.fr
Système universitaire de documentation (SUDOC)	www.sudoc.abes.fr

Sites américains et internationaux	
Bibliothèque du Congrès (Washington)	www.loc.gov.
Dublin Core	http://dublincore.org
International Federation of Library Association and Institutions (IFLA)	www.ifla.org
Online Computer Library Center (OCLC)	www.oclc.org
Unicode	www.unicode.org

Sites canadiens	
Bibliothèque nationale du Canada	www.nlc.bnc.ca
Hapax	http://iquebec.ifrance.com/hapax

Sites russes	
Bibliothèque d'État de Russie	www.rsl.ru
SONEGOS	www.rsl.ru/SONEGOS
Welcome to The Home of KOI8-R-Russian Net Character Set.	http://koi8.pp.ru

ANNEXES

Annexe 1 : Comment installer la prise en charge d'Unicode sous Windows 98 et NT4 ?

Extrait de : Documentation technique sur l'Unicode par Vincent Colnat (47)

I. Système compatible avec Unicode : Windows 2000, Windows XP

Pour installer Unicode il faut utiliser le programme : Ajout/ Suppression de programmes disponible dans *Démarrer / Paramètres / Panneau de Configuration.*

Ensuite dans l'option *Modifier /Supprimer des programmes,* sélectionner Miscrosoft Office 2000 CD-ROM2 (Disponible sous Office XP) et cliquer sur modifier.
Ensuite cliquer avec le bouton gauche sur Police de caractères universelle disponible sous le chemin : *Microsoft Office / Outils Office / Polices /Police de caractères universelle*. Valider Exécuter à partir du disque dur.

II. *Installer la prise en charge de Unicode sous Windows 98 et NT 4*

Ces OS gère Unicode mais l'encodage n'est pas en 16 bits.
Pour afficher une langue spécifique sous Windows 98, il vous suffit d'installer le sous-programme *"prise en charge de la langue spécifique "*.
Ce sous-programme est inclus dans *"Explorer"*(Oui ! ce n'est pas logique, mais c'est comme ça !) Ainsi, **même si vous utilisez un autre navigateur ou un autre programme**, il vous suffit d'installer *"Explorer"* avec l'option *"prise en charge de la langue spécifique "* pour Windows 98 ou NT 4.
Vous avez besoin de télécharger des composants à partir du site *"Explorer"* ou d'un CD-Rom.
Trouver et lancer un fichier *"ieXsetup.exe"* (si vous n'en disposez pas, alors vous le téléchargez à partir de Microsoft Téléchargement.
Choisir l'option d'installation "*Installation personnalisée*" (**c'est très important** et il ne faut pas effectuer *"installation par défaut"*)
Choisir les options *"Sélection automatique de la langue"* et *"prise en* charge de la langue spécifique ".
Internet Explorer 5.x et suivants
Installer. Vous avez besoin de télécharger des composants à partir du site *"Explorer"* ou d'un CD-Rom.
Trouver et lancer un fichier *"ie5setup.exe"* (Si vous n'en disposez pas, alors vous le téléchargez à partir de Microsoft Téléchargement.
Choisir l'option d'installation "Installation personnalisée" (**c'est très important** et il ne faut pas effectuer "installation par défaut")
Choisir les options "Sélection automatique de la langue" et "prise en charge de la langue désirée".
• Configurer.
• Menu "Affichage"

• Sous-menu *"Codage"*
• Cliquer sur *"Plus"*
• Cliquer sur *"Unicode (UTF-8)"*. Si *"Unicode (UTF-8)"* n'apparaît pas, installer le langage avec Windows 98/NT4 ou Windows 2000 , Windows XP.
Vous devez avoir dans le sous-menu *"Codage"* **au moins** *"Unicode (UTF-8)"*, *"Alphabet occidental (Windows)"*.
• Cocher la case *"Sélection automatique"*. Pour plus de facilité

Netscape Communicator 6.x
Menu *"Afficher"*
Sous-menu *"Codage des caractères"*
Sous sous-menu *"Autres"*
Choisir *" Unicode (UTF 8)"*.
Revenir à *"Afficher"*, *"Codage des caractères"*
Sous sous-menu *"Détection auto"*
Choisir "*Détection automatique (tout)"*.

Office 2000/XP (Word, Excel, Access)
Vous avez besoin du cd-rom d'installation de votre programme.
1 "Panneau de configuration"
2 "Ajout/Suppression de programmes"
"Microsoft Office" (ou Word, Excel, Access...)
• La case *"Modifier"*
• La case *"Ajouter/Supprimer des composants"*
• Développer le répertoire *"Outils Office""*
• Sélectionner le sous répertoire *"Outil Paramètres linguistiques""*
• Sélectionner *"Exécutez à partir du disque""*
• La case *"Mise à jour"*.
Configurer. Spécifier à votre programme que vous chargez des pages dans une langue spécifique
• Menu *"Démarrer"*
• Sous-menu *"Outils Microsoft Office"*
"Paramètres linguistiques Office""
Dans la liste "*Afficher les commandes et activer l'édition de*", cocher la case correspondant à la langue spécifique.
Vous pouvez (optionnel) dans Word 2000 choisir de différentier les caractères latins et les caractères Grecs par une police, une taille, un effet.
• Menu *"Format"*
• Sous-menu *"Polices..."*
• Onglet *"Polices, styles et attributs"*
Il y a 2 zones *"Police de caractères latins"* et juste en dessous *"scripte complexe"*.
C'est là que vous faites varier les caractéristiques.
Les caractéristiques deviennent permanentes quand vous cliquez tout en bas *"par défaut"*.

Annexe 2 : Tableau de translittération pour les caractères cyrilliques russes

Caractère cyrillique	Translittération en caractères latins
А	A
Б	B
В	V
Г	G
Д	D
Е	E
Ё	Ë
Ж	Ž
З	Z
И	I
Й	J
К	K
Л	L
М	M
Н	N
О	O
П	P
Р	R
С	S
Т	T
У	U
Ф	F
Х	H
Ц	C
Ч	Č
Ш	Š
Щ	Ŝ
Ъ	”
Ы	Y
Ь	‘
Э	È
Ю	Û
Я	Â

Source : Norme française ISO 9 (Juin 1995) / AFNOR 1995

Annexe 3 : Comment saisir à l'aide de son clavier français, les lettres diacritées Č, Š, Ž, et Ŝ ?

- Dans Word (Windows 97), une solution consiste à personnaliser son clavier pour saisir ces lettres facilement.

Il suffit de faire l'opération suivante :

- Aller dans le menu de Word
- Choisir l'onglet Insertion
- Puis Caractères spéciaux (Latin étendus A)
- Sélectionner les lettres voulues
- Cliquer sur touches de raccourci
- Choisir la commande pour chaque lettre ("attribuer").

Les lettres manquantes sur du clavier français, mais néanmoins nécessaires pour la translittération des caractères cyrilliques (langue russe) n'étant qu'au nombre de quatre, la mémorisation des commandes est possible.

Cette solution ne permet néanmoins pas d'afficher les lettres manquantes dans un fichier Excel 97 SR-2. Il faut donc trouver une autre solution.

- Dans un fichier Excel 97 SR-2 :

Dans le menu "Démarrer" de Windows, aller dans "Accessoires" puis choisir l'onglet "Table des caractères Unicode". Cette table permet de stocker dans le presse-papier les caractères dont nous souhaitons disposer.
Cette solution n'est pas ergonomique mais elle permet d'obtenir les quatre caractères diacrités dans un fichier Excel 97 SR-2. Elle a l'avantage de pouvoir respecter scrupuleusement la norme ISO.

- Dans Windows Millenium :

Sous Windows Millenium, il suffit de cliquer sur « Caractères spéciaux » dans le menu « Insertion », de cliquer ensuite sur l'onglet « Symboles », de choisir Arial Unicode MS et de créer des touches de raccourci pour les quatre lettres.

Cette opération permet la saisie des quatre caractères dans des fichiers Word et Excel, mais pas dans l'interface d'un navigateur.

Annexe 4 : Récapitulation des réponses des éditeurs de logiciels

GESTION DES CARACTERES LATINS ETENDUS ET DES CARACTERES NON LATINS	
Votre SIGB est-il censé pouvoir manipuler des données en caractères latins étendus et des caractères non latins?	Ever Team : Oui ExLibris : Oui Éditeur XX : notre SIGB gère les caractères latins et non latins (grec, cyrillique) mais ne gère pas dans la version actuelle les caractères arabes et asiatiques Éditeur YY : Oui, notre SIGB gère les jeux de caractères prévus dans Unimarc dont ISO 54.26. Le produit est installé en standard avec ces normes
Avez-vous déjà testé cette possibilité ?	Ever Team : Oui ExLibris : Oui Éditeur XX : Oui Éditeur YY : Oui

POLICES, CLAVIERS	
Pour utiliser ces polices de caractères, des claviers matériels spécifiques sont-ils fournis ?	Ever Team : Non ExLibris : Non Éditeur XX : Oui, un clavier est disponible sur tous les écrans de saisie (catalogage, inscription lecteurs, etc.) Éditeur YY : Nous ne fournissons pas de clavier spécifique. En catalogage, les caractères spécifiques ou diacritiques, hormis ceux qui sont sur les claviers standard ASCII, sont accessibles via une fenêtre ou par raccourcis clavier. Les bibliothèques peuvent s'équiper de claviers spécifiques si elles le souhaitent.

Si la saisie de tous les caractères sur clavier n'est pas possible ou si aucun logiciel de saisie n'est fourni, quels moyens envisagez-vous ?	Ever Team : Pour gérer les cas particuliers comme le Chinois, le Japonais, le Coréen, deux solutions sont envisageables : o Utilisation des fonctions de copier/coller (à partir de la fonction insertion de caractères spéciaux de solutions externes) o Utilisation de solutions tiers de traduction à la volée (Twin Bridge) ExLibris : Existence de claviers virtuels (paramétrés dans Aleph) ou utilisation de claviers IME du marché. Éditeur XX : - Éditeur YY : -

STOCKAGE DES DONNEES	
Format de stockage des données (bibliographiques) ?	Ever Team : Unimarc, LCMarc, USMarc, Marc 21, Intermarc . Loris est une solution indépendante des formats MARC, permettant d'envisager un stockage des données sous d'autres formats. ExLibris : Unicode UTF-8 Éditeur XX : Format MARC (Unimarc, Marc21), Fabritius (basé sur les recommandations de Paul Getty), Dublin Core, Smart (créé par Geac) Éditeur YY : Unimarc et OCLC-Marc, en France.
Ce format est-il globalement compatible avec Unicode ?	Ever Team : Oui ExLibris : - Éditeur XX : Oui Éditeur YY : Oui
Codage de stockage des données ?	Ever Team : UTF-8 ExLibris : Unicode UTF-8 Éditeur XX :- Éditeur YY : Unimarc et identifiants Unicode si le caractère n'est pas dans le jeu Unimarc.

SAISIE / CATALOGAGE	
UNIMARC prend-il en compte les caractères non latins ? Si, oui, comment ?	Ever Team : En fonction des normes de codages adoptées, Loris est à même de s'adapter à la norme. ExLibris : Dans le label (champ Unimarc) sont stockés la langue de catalogage, l'alphabet et le codage utilisés. Aleph ajoute pour chaque champ l'alphabet du champ : latin, arabe, hébreu, grec, cyrillique, CJK (chinois, japonais, coréen). Éditeur XX : Oui, clavier fourni. Éditeur YY : Les données sont enregistrées en unimarc qui prend en compte plusieurs alphabets : grec, cyrillique … (voir nouvelle édition d'Unimarc)/
Est-il possible d'intégrer des procédures de conversion d'une écriture dans l'autre pour ne faire qu'une saisie ?	Ever Team : Oui, fonction non proposée en standard, mais pouvant être mis en œuvre par un simple paramétrage. ExLibris : C'est possible pour certaines langues : translittération arabe vers arabe (langue d'origine) par exemple. Éditeur XX :- Éditeur YY : Non
Les différents sens d'écriture peuvent-ils être pris en compte par le SIGB ?	Ever Team : Oui, gestion prise en compte par l'interface. ExLibris : Oui. http://libnet1.ac.il/~libnet/uls/uls.htm - Choisir ULS Web catalog – Keywords anywhere – Saisir arabic → Liste de titres arabes, hébreux, latins Éditeur XX :- Éditeur YY : Oui.
Est-il possible de créer des bases de données dans toutes les écritures, permettant de stocker ces données et d'importer des informations dans toutes les écritures ?	Ever Team : Oui, toutes écritures reconnues par UNICODE peuvent être gérés par Loris. (pour information : un certain nombre de caractères chinois définis dans MARC 21 n'existe pas encore dans UNICODE). ExLibris : Oui Éditeur XX :- Éditeur YY : Oui avec Unicode qui est géré au niveau de la base de données.

Possibilité de faire des notices d'autorité recensant : - toutes les formes du titre ? - les graphies des noms d'auteurs en caractères originaux et latins, quelle que soit l'écriture utilisée à l'interrogation ?	Ever Team : Oui ExLibris : Oui Éditeur XX : Oui, à confirmer. Éditeur YY : Notices d'autorités Unimarc : Auteurs/Collectivités, Matières, Titres de collections, Titres uniformes avec gestion des renvois VOIR et VOIR AUSSI. Ever Team : Oui, lors d'une recherche Loris répondra quelle que soit la forme de l'interrogation. ExLibris : Oui Éditeur XX :- Éditeur YY : Oui.

INDEXATION	
Votre SIGB indexe-t-il les caractères non latins ?	Ever Team : Oui ExLibris : Oui Éditeur XX : ne sait pas. Éditeur YY : Oui.
Possibilité d'indexer toutes les formes du titre (dans le cas de traductions), c'est-à-dire en caractères originaux et latins ? Et sur la base de quels alphabets ? Est-il possible d'avoir cette présentation : « titre en langue originale » = « titre du livre dans la langue du livre »	Ever Team : L'indexation est réalisée quelle que soit la langue avec un tri implicite en fonction de critères personnalisables. La norme UNICODE permet à tout moment (indexation, stockage, etc…) de différencier les langues. ExLibris : Oui. Latin, arabe, hébreu, grec, cyrillique, CJK (chinois, japonais, coréen). Éditeur XX :- Éditeur YY : Oui. Tous les alphabets gérés par Unicode. Ever Team : Oui, par simple paramétrage. ExLibris : Oui. Éditeur XX : - Éditeur YY : Oui.

Création possible de tout index ou fichier inversé dans toutes les langues ?	Ever Team : Oui, (par exemple si le gestionnaire désire indexer le champ 245a, les deux langues respectivement sur champ 245a seront indexées). ExLibris : Oui Éditeur XX : Oui Éditeur YY : Oui

RECHERCHE	
Tous les modes de recherche (rapide, plein texte, mode assisté, mode expert, ….) sont-ils envisageables ?	Ever Team : Oui, Loris propose en standard tous ces modes de recherche. ExLibris : Oui Éditeur XX : Oui Éditeur YY : Oui. Interrogation en texte intégral sur la base de données de documents numérisés (Horizon, Bibliothèque électronique) avec un moteur de recherche sémantique (RetrievalWare). Au niveau du catalogue, possibilité de créer un index sur toutes les zones de la notice bibliographique.
Possibilité d'insérer des thésaurus multilingues en caractères de tous types ? Dans quelles limites ? (a priori, combinaisons infinies)	Ever Team : Oui, pas de limites identifiée. EVER TEAM traite par exemple le thésaurus EUROVOC (11 langues, UTF-8) ExLibris : Oui Pas de limites connues. La communauté européenne utilise des thésaurus sur 10 langues. Éditeur XX :- Éditeur YY : Non au niveau de l'OPAC (mais possible avec le moteur RetriewalWare utilisé pour les bases de documents électroniques).

Techniques de recherche : ♦ Interrogation en texte intégral : possible dans toutes les langues ?	Ever Team : Oui, cette fonctionnalité est en cours de validation ExLibris : Oui. Texte intégral de la notice. Éditeur XX : Dans les caractères gérés par le système. Éditeur YY : Voir commentaire ci-dessus sur le moteur RetrievalWare. Il gère plusieurs langues : FR, ENG, GER, SPA, IT
♦ Votre moteur de recherche est-il utilisable pour tous les jeux de caractères ?	Ever Team : - ExLibris : Oui Éditeur XX : ISO 5426, ISO 646 Éditeur YY : -
♦ Lequel est-ce ?	Ever Team : - ExLibris : Moteur propriétaire CCL (common command language) Éditeur XX :- Éditeur YY : RetrievalWare de la société Convera
♦ Dictionnaires multilingues ?	Ever Team : Oui, disponible sous forme d'assistant multilingue (de type index ou thesaurus) ExLibris : Oui pour le chinois / pinyin Éditeur XX : Oui. Éditeur YY : Non pour l'OPAC mais possible avec RetrievalWare
Les techniques linguistiques utilisées sont-elles adaptées à des recherches multilingues ?	Ever Team : Oui. ExLibris : Oui Éditeur XX : Oui. Éditeur YY : On ne gère pas d'outils linguistiques au niveau de l'OPAC
Possibilité d'interroger dans une langue et de trouver des documents écrits dans d'autres langues ?	Ever Team : Oui. Par exemple, au niveau des autorités, des thesaurus multilingues, etc… ExLibris : Oui, si utilisation du thésaurus multilingue. Éditeur XX : Oui. Éditeur YY : Oui, si la notice est multilingue.

OCR	
Océrisation possible pour toutes les écritures ? Quel est votre logiciel OCR ?	Ever Team : Non, Loris s'appuie sur la solution d'OCR FineReader professionnal. Cette solution leader est capable d'identifier 122 langues dans sa version européenne et 177 langues dans sa version cyrillique plus. ExLibris : Aleph est un SIGB. Il ne comporte pas d'OCR. Il peut être couplé avec notre outil GED intitulé « DigiTool ». Celui-ci permet de stocker/consulter des objets numériques de différents types (texte, image, vidéo, audio) et de leur attribuer des métadonnées (notice de catalogage, données techniques et copyright). DigiTool permet de créer des liens hiérarchiques entre ces objets et de réaliser des recherches en texte intégral sur les documents de type texte. Éditeur XX : Inexistant. Éditeur YY : Oui, avec notre SIGB, c'est une option. Éditeur YY : LightScanning qui inclut un OCR

AFFICHAGE	
Affichage des résultats en caractères romanisés, diacrités, et en caractères originaux ?	Ever Team : Oui, pour les trois types d'affichage cités. ExLibris : Oui Éditeur XX : Oui. Éditeur YY : Non, il n'y a pas de gestion de correspondances entre les différentes formes.
Possibilité d'afficher les références de façon à ce qu'elles précisent si les documents sont en langue originale, traduits et dans quelles langues, bilingues ?	Ever Team : Oui. L'affichage des références résulte directement du catalogage effectué par les professionnels. ExLibris : Oui Éditeur XX : Oui. Éditeur YY : Zone codée dans Unimarc pour indiquer la langue originale et la langue du document

PRODUCTION ET CONSULTATION DES DONNEES EN CARACTERES LATINS ETENDUS ET EN CARACTERES NON LATINS	
Votre SIGB est-il livré avec des fontes appropriées ?	Ever Team : En cas de lacunes sur le poste Client certaines fontes peuvent être proposées Arial UNICODE (pratiquement toutes les langues), Bilstream cybertbit ou Code 2000 (pour gérer le Birman, etc…) ExLibris : Oui pour le serveur. Les clients utilisent des fontes Windows. Utiliser 2000 ou XP pour avoir plus de choix. Éditeur XX :- Éditeur YY : Traitement de texte intégré à l'application Horizon. Les polices Unicode sont livrées en standard.
De quel type sont-elles (TrueType, PostScript, OpenType, autre) ?	Ever Team : les fontes proposées disposent des caractères cités (TrueType, PostScript, OpenType, etc…) (ces critères étant indépendants de la gestion du multilinguisme) ExLibris : Fontes Windows standard Éditeur XX :- Éditeur YY : True Type.
Sont-elles fournies réellement pour toutes les langues annoncées ?	Ever Team : Oui. ExLibris : Fontes Windows standard Éditeur XX :- Éditeur YY : Oui

FORMAT D'ÉCHANGE	
Importation possible de fichiers / notices / listes d'autorité / en langues étrangères (caractères non latins) et dont les formes de catalogage sont différentes ?	Ever Team : Oui. Concrètement, on rencontre 2 cas : 1. Format MARC : o USMARC : utilisation du jeu de caractères ANSEL (reconnu également par MARC 21) o UNIMARC : utilisation du jeu de caractères ISO 5426 (pour les caractères accentués) 2. Autres formats : Loris possède des convertisseurs génériques au format UNICODE. ExLibris : Oui. Format ISO 2709 (Marc 21, MAB, Unimarc) en standard. Éditeur XX : Oui. Éditeur YY : Pour les notices en format MARC, fonction d'import/export ISO 27.09 fourni en standard.
Que se passe-t-il lors de l'importation de notices étrangères qui ne sont pas au format MARC / ?	Ever Team : Cf. ci-dessus. ExLibris : Il faut créer des programmes de conversion spécifiques. Éditeur XX : Outils de Conversion inclus dans l'import. Éditeur YY : Conversion des notices en Unimarc

GESTION DES PRETS	
Quel est le logiciel utilisé par la messagerie électronique ?	Ever Team : Loris gère directement la messagerie système. ExLibris : Logiciel intégré dans Aleph, utilisant le serveur méls de la bibliothèque. Éditeur XX : Oui. Éditeur YY : La messagerie n'est pas propre au produit. Utilisation de la messagerie installée.
Permet-il la communication dans toutes les langues avec le client ? (relances, réservations dans n'importe quelle langue, par messagerie Internet,….)	Ever Team : Oui. ExLibris : Oui. Le code langue de communication est stocké dans l'enregistrement du lecteur, du fournisseur. Éditeur XX : Oui Éditeur YY : Non. Des évolutions sont prévues (avis aux usagers par exemple).

INTERFACE WEB	
Compatible avec tous les jeux de caractères ?	Ever Team : Oui. ExLibris : Les navigateurs utilisés doivent comprendre Unicode (Netscape, Mozilla, IE…) Éditeur XX : Idem que pour les notices bibliographiques Éditeur YY : Unicode.
Possibilité de créer sur une notice des liens hypertextes avec le document afin d'avoir à l'écran l'image d'une page ou plusieurs du document (page de titre et du sommaire) ?	Ever Team : Oui, Loris permet de gérer tous types de liens : champs MARC (exemple UNIMARC : (Bloc 4XX, champ 856 (URL), etc.), module GED, etc.) ExLibris : Oui (champ Unimarc 856). Éditeur XX : Oui. Éditeur YY : Via la zone structurée multimedia 856. Il est aussi possible de créer des liens URL sur les autres zones bibliographiques ainsi que sur celles de la notice d'exemplaire.

PORTAIL	
Comment le client étranger peut-il personnaliser ses profils dans une langue étrangère ?	Ever Team : Quelle que soit la langue utilisée par le client étranger, celui-ci pourra définir ses profils dans sa langue d'origine. En standard, le système lui retournera toutes données du système répondant à la requête quelle que soit la langue. ExLibris : L'interface de l'OPAC est multilingue. La langue de l'interface est définie sur le profil de l'utilisateur (identifié ou invité). Il peut la modifier. Éditeur XX :- Éditeur YY : Oui, possibilité de proposer plusieurs interfaces de dialogue en différentes langues.
Peut-on définir pour chaque profil la langue utilisée pour la diffusion de l'information ?	Ever Team : Oui, en standard de gestion de l'affichage multilingue. ExLibris : Oui – Dépend de la langue définie sur l'enregistrement du lecteur. Éditeur XX : Oui, par défaut 3 langues sont fournies en standard (français, anglais, néerlandais) ; possibilité pour la bibliothèque d'ajouter d'autres langues d'interrogation. Éditeur YY : non.

AIDE EN LIGNE	
Le client étranger peut-il bénéficier d'une aide en ligne, dans sa langue ?	Ever Team : Oui, l'aide en ligne web peut tout à fait être étendue afin de proposer au client étranger une aide en ligne dans sa langue. ExLibris : Oui Éditeur XX : Oui. Éditeur YY : Oui s'il participe à la traduction, hormis pour les langues où le produit est déjà distribué.

SÉCURITÉ	
Peut-on envisager des mots de passe en caractères non latins ?	Ever Team : Oui. ExLibris : Oui Éditeur XX : A étudier. Éditeur YY : Non

Annexe 5 : Liste non exhaustive de codages informatiques de caractères non latins

Année	NOM DU CODAGE		Nombre de caractères	LANGUES / ECRITURES
1968	**ASCII sur 7 bits**	ou US-ASCII ou ISO 646 ou ANSI X3.4	128	Langues latines non accentuées (anglais)
	KOI			
		KOI-8 Russe KOI-8 Ukrainien		Norme soviétique
Seconde moitié des années 1980	**ISO 8859-N** (avec 1 < N <15)		256	Cyrillique, grec, hébreu, arabe, thaï
		ISO 8859-1	256	tous les alphabets d'Europe occidentale
		ISO 8859-2	256	tous les alphabets d'Europe centrale (sauf cyrillique)
		ISO 8859-3	256	maltais, turc, esperanto
		ISO 8859-4	256	langues pays baltes, lapon, groëlandais
		ISO 8859-5	256	alphabet cyrillique
		ISO 8859-6	256	arabe
		etc		
		ISO 8859-15 (adaptation de l'ISO 8859-1)	256	permet le "e dans l'o" et comporte le symbole de l'euro
	ISO 5426, 5426-2		256	Jeux de caractères latins, latins rares, cyrilliques, grecs pour l'échange de notices bibliographiques (format UNIMARC)

Année	NOM DU CODAGE	Nombre de caractères	LANGUES / ECRITURES
	Les jeux étendus Windows		
	CP 1252 ou Win Latin 1		Jeu utilisé par les PC sous Windows NT pour le français
	CP 1251 ou WinCyrillic		Il ne s'agit pas de la norme ISO 8859-5 avec des caractères étendus. Elle est donc incompatible avec ISO 8859-5 et avec la norme soviétique KOI-8
1990	**Unicode**	65 536	Jeu de caractères universel
1993	**ISO 10 646**		

Sources :

- Codage informatique des systèmes d'écriture Version 2.0.1 / Gianna Vacca
  http://giannieanna.chez.tiscali.fr/codagecharsets201.pdf
  (site consulté en février 2003)
- Le traitement informatique des documents en caractères non latins / A. Dupas.- Tour d'horizon. – BBF, Paris,T.43, n°1, 1998
  http://bbf.enssib.fr/bbf/html/1998_43_1/1998-1-p104-dupas.xml.asp
  (site consulté en février 2003)

Annexe 6 : Bibliothèques et formats de catalogage

Bibliothèque	Logiciel de gestion de bibliothèque	Formats de catalogage
BDIC	- ALEPH pour les documents en russe et en bulgare depuis 1997. - GEAC pour les catalogues en ligne (notices en caractères latins)	
BnF		INTERMARC pour bases BN-OPALE PLUS
Bibliographie nationale française sur CD-Rom		UNIMARC
FNSP		INTERMARC
Ecole Polytechnique		US MARC
Library of Congress		USMARC (MARC 21, à présent)
Réseau des bibliothèques de l'ex-URSS : SONEGOS Bibliothèque d'Etat de la Russie (RGB)	Logiciel ALEPH	USMARC

Sources :

- Unimarc, manuel de catalogage / M-R Cazabon. – Edition du Cercle de la Librairie, 1993
- www.bdic.fr/catalogues.htm
- IFLA Journal 28 (2002) 2
  The SONEGOS Website as a Gateway to the Libraries of the Commonwealth of Independent States / M. Shvarstman, (p.69-73)

Notice

Description bibliographique :
Constitution d'un annuaire de chercheurs internationaux et catalogage dans un SIGB : les problèmes liés à l'intégration d'informations multilingues. / Hélène Rousselot. – 2003. – 95 p. – Mémoire DESS : information-documentation : Paris, INTD : 2003.

Mots-clés :
Caractères cyrilliques, Catalogage, Clavier virtuel, Codage informatique, Internet, MIME, Moteur de recherche, Multilinguisme, Romanisation, Système intégré de gestion de bibliothèque, Unicode.

Résumé :
L'alimentation d'un Annuaire de chercheurs internationaux sous la base Access, à partir de ressources de la Toile et l'étude d'une grille d'analyse, remplie par des éditeurs de SIGB, permettent de comprendre le rôle déjà joué par Unicode dans le fonctionnement d'un système informatique pour le codage des caractères non latins.
L'examen de requêtes effectuées sur deux catalogues en ligne montre les difficultés posées par la romanisation dans les notices bibliographiques et d'autorité. Certaines de ces questions peuvent être réglées par l'adoption d'Unicode qui permet l'introduction des caractères originaux dans les notices.
Un groupe d'experts, piloté par D. Duclos-Faure, fournit des recommandations pouvant s'appliquer aux centres de documentation et aux bibliothèques pour la gestion de leurs fonds en caractères non latins.
Au-delà de ces questions techniques relevant du domaine informatique et du catalogage, le rôle clé que jouera Unicode dans le développement de la production, du traitement et de la diffusion de l'information en langues non latines est un point de départ pour des réflexions quant aux enjeux culturels et politiques de la production et de la circulation de l'information dans des langues autres que l'anglais.