Gouvernement en ligne

VERSION ENRICHIE DU STANDARD SUR LES JEUX DE CARACTÈRES CODÉS (SGQRI 003)

Version 1.0 du 11 décembre 2006

Québec

Table des matières

Remarque :
Élaboré par le ministère des Services gouvernementaux, le standard adopté par le Conseil du trésor le 11 décembre 2006 se trouve dans le *Recueil des politiques de gestion* du Conseil du trésor (RPG 11 2 2 9). Ce document porte exclusivement sur les éléments obligatoires pour les ministères et les organismes.

Afin de faciliter la mise en place de ce standard dans l'Administration gouvernementale, le ministère des Services gouvernementaux rend disponible une version enrichie, à la manière d'une version annotée, dans le Recueil des éléments normatifs en matière de ressources informationnelles. Ce document reprend le contenu du standard adopté par le Conseil du trésor en y ajoutant des renseignements supplémentaires pertinents et d'autres éléments qui ne sont pas de nature obligatoire.

Les renseignements supplémentaires dans cette version enrichie sont présentés dans des encadrés en grisé et peuvent porter, notamment sur une mise en contexte, un exemple, une recommandation, une remarque, une déclaration sur la conformité ou sur la dérogation de ce standard à d'autres standards du gouvernement du Québec, ou une déclaration sur la conformité au concept d'adaptabilité culturelle et linguistique. Ils indiquent finalement la composition du groupe de travail responsable de l'élaboration du standard.

VERSION ENRICHIE DU STANDARD SUR LES JEUX DE CARACTÈRES CODÉS (SGQRI 003)

SECTION I : DISPOSITIONS GÉNÉRALES

S.-s. 1 – Objet du standard

1. Ce standard détermine les jeux de caractères codés qui peuvent être utilisés, et ce, dans le but de faciliter l'interopérabilité de ces jeux.

Remarque :
La version enrichie du standard aborde aussi certaines considérations de base concernant le codage des caractères dans Internet.

S.-s. 2 – Champ d'application

2. Ce standard s'applique aux ministères et aux organismes visés par l'article 64 de la Loi sur l'administration publique (L.R.Q., c. A-6.01).

Loi sur l'administration publique (L.R.Q., c. A-6.01) :

CHAPITRE I
OBJET ET APPLICATION
Composition.

3. Pour l'application de la présente loi, l'Administration gouvernementale est constituée :

1° des ministères du gouvernement ;
2° des organismes budgétaires, soit les organismes dont tout ou partie des dépenses sont prévues aux crédits qui apparaissent dans le budget de dépenses déposé à l'Assemblée nationale sous un titre autre qu'un crédit de transfert ;
3° des organismes dont le personnel est nommé suivant la Loi sur la fonction publique (chapitre F-3.1.1) ;
4° des organismes dont le gouvernement ou un ministre nomme la majorité des membres ou des administrateurs et dont au moins la moitié des dépenses sont assumées directement ou indirectement par le fonds consolidé du revenu.

Organisme.
Est considérée comme un organisme, une personne nommée ou désignée par le gouvernement ou par un ministre, avec le personnel qu'elle dirige, dans le cadre des fonctions qui lui sont attribuées par la loi, le gouvernement ou le ministre.

Applicabilité.

4. L'Assemblée nationale, toute personne nommée ou désignée par cette dernière pour exercer une fonction en relevant, avec le personnel qu'elle dirige, ainsi que la Commission de la

représentation ne sont assujetties à la présente loi que dans la mesure prévue par une loi.

Il en est de même des tribunaux au sens de la Loi sur les tribunaux judiciaires (chapitre T-16), des organismes dont l'ensemble des membres sont juges de la Cour du Québec, du Conseil de la magistrature et du comité de la rémunération des juges de la Cour du Québec et des cours municipales.

CHAPITRE VI
GESTION DES RESSOURCES INFORMATIONNELLES
Application.
64. Le présent chapitre s'applique à l'Administration gouvernementale.

S.-s. 3 – Définitions

3. Dans le présent standard, on entend par :

a) **caractère** : un élément d'un ensemble utilisé pour organiser, commander ou représenter des données ;

(Source : norme ISO/CEI 10646 *2003* [Jeu universel de caractères codés sur plusieurs octets {JUC}] de l'Organisation internationale de normalisation [ISO] et de la Commission électrotechnique internationale [CEI])

Remarque :
Les caractères sont principalement, mais pas exclusivement, les lettres, la ponctuation et les autres signes qui sont utilisés dans les notations techniques ou les textes en langue naturelle.

b) **caractère codé** : un caractère et sa représentation numérique ;

Remarque :
Cette définition exclut les images tramées (*bitmaps*) représentant du texte et les bandes sonores ou vidéo numérisées, mais inclut le texte dans les formats d'image où le texte est codé ainsi que les sous-titres formés de caractères codés dans certains formats vidéo.

c) **jeu de caractères codés** : un ensemble de règles univoques qui définissent un groupe de caractères et établissent une correspondance entre chaque caractère et sa représentation codée ;

d) **répertoire** : un ensemble précis de caractères, défini de manière indépendante du codage.

Remarque :
Le répertoire d'un jeu de caractères codés est la liste des caractères visibles que ce jeu peut représenter. La manière la plus précise de décrire ces caractères est par leur nom, et non par leur représentation, qui peut être très variable et porter à confusion. Il existe une liste normalisée et exhaustive des caractères du monde en français. Elle se trouve dans la version française de la norme internationale ISO/CEI 10646.

SECTION II : SPÉCIFICATIONS

S.-s. 1 – Conditions de conformité au standard

4. Un jeu de caractères codés utilisé dans une application informatique est conforme au présent standard s'il respecte les exigences de la sous-section 2.

S.-s. 2 – Exigences

Principes d'application :
Ce standard s'applique à toute partie de l'infrastructure gouvernementale qui a à traiter des données textuelles sous quelque forme que ce soit. Ces données comprennent, notamment les documents contenant du texte, les bases de données incluant des champs de nature textuelle, les messages électroniques (courriels, messages publiés dans des forums, etc.), les métadonnées telles que les adresses, noms de fichiers et noms de domaine, les éléments textuels des interfaces personne-machine y compris les messages d'erreur, les formulaires électroniques, etc.

Une grande proportion des données stockées et traitées par les systèmes informatiques est constituée de texte. Pour assurer son traitement informatique, un texte doit être codé. Or il existe un grand nombre de codages, inventés au cours des cinq dernières décennies, qui ont des propriétés et des capacités différentes. Pour les besoins de l'infrastructure gouvernementale, il importe de choisir et de désigner certains codages en vertu de trois critères principaux :

- tous les codages choisis doivent permettre de représenter tous les caractères nécessaires et, dans la mesure du possible, être utiles au codage de la langue française ;

- il importe de minimiser le nombre de codages différents choisis dans le but de faciliter les échanges et de diminuer les risques d'erreurs et de pertes de données causés par la présence de plusieurs codages ;

- il faut tenir compte dans une certaine mesure du contexte, de l'offre du marché et de son évolution prévisible.

Il va de soi que pour que les échanges soient possibles entre jeux de caractères, dans une opération appelée transcodage, un dénominateur commun doit exister. C'est ici qu'intervient la notion de *répertoire*.

L'ensemble des caractères d'un jeu de caractères donné, défini de manière indépendante du codage, est appelé dans le jargon de ce domaine de normalisation un *répertoire*. Dans le monde de la normalisation des jeux de caractères codés, il est aujourd'hui reconnu, compte tenu que même la représentation visible d'un caractère est insuffisante pour identifier un caractère (plusieurs caractères codés différents, parfois à l'intérieur d'un même jeu de caractères, peuvent de nos jours partager la même représentation et ne pas être repérables dans une base de données avec les mêmes paramètres de recherche[1]) que c'est le nom de ce caractère, dans

[1] Par exemple, la lettre majuscule latine A, la lettre majuscule grecque ALPHA et la lettre majuscule cyrillique A (utilisée en russe et en ukrainien, notamment) partagent toutes exactement la même représentation dans une police

une langue donnée, qui permet le mieux de déterminer un caractère codé. Quand deux noms officiels de caractères sont identiques entre deux jeux de caractères codés, il est possible de prétendre qu'il y a équivalence. Il faut alors retenir que le nom exact d'un caractère revêt une importance technique primordiale.

Remarque :
La figure qui suit illustre la distinction entre un répertoire (une liste de caractères définie de manière indépendante du codage – définie par les noms de caractères), qui peut être commun à plusieurs jeux, et un jeu précis de caractères codés (tableau définissant la correspondance entre une valeur codée et sa représentation).

	Jeu de caractères codés de la norme internationale ISO/CEI 8859-15 (Alphabet latin n° 9)	
Répertoire	**Caractère codé**	**Représentation**
LETTRE MAJUSCULE LATINE A	65	A
LETTRE MAJUSCULE LATINE B	66	B
LETTRE MAJUSCULE LATINE C	67	C
POINT D'INTERROGATION	63	?
POINT D'EXCLAMATION	33	!
SYMBOLE EURO	164	€
…		

Répertoire essentiel du français :
Voici la liste des caractères considérés comme essentiels au soutien du français au sein de l'Administration gouvernementale.

!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
0	1	2	3	4	5	6	7	8	9	:	;	<	=	>
?	@	A	B	C	D	E	F	G	H	I	J	K	L	M
N	O	P	Q	R	S	T	U	V	W	X	Y	Z	[	\
]	^	_	`	a	b	c	d	e	f	g	h	i	j	k
l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
{	|	}	~	«	»	À	Â	Ä	Æ	Ç	È	É	Ê	Ë
Î	Ï	Ù	Û	Ü	à	â	æ	ç	è	é	ê	ë	î	ï
Ù	û	ü	ÿ	Ÿ	œ	Œ								

Cette liste reprend certains caractères qu'il est pour le moins douteux de considérer comme essentiels en français (ex. : /, #, \) ; ils sont néanmoins utiles, notamment dans les systèmes informatiques. Aussi, la lettre majuscule Ÿ – malgré la présence de la minuscule ÿ – et les digrammes œ et Œ, pourtant essentiels à l'orthographe du français sont absents de l'un des jeux de caractères les plus importants qui soient, celui de la norme ISO/CEI 8859-1 (Traitement de l'information – Jeux de caractères graphiques codés sur un seul octet, Partie 1 : Alphabet latin n° 1) de l'ISO et de la CEI. Cette situation comporte cependant aujourd'hui des solutions vers lesquelles nous devons tendre, puisque le but est de soutenir intégralement le français et

de caractères donnée. Ces trois caractères peuvent tous (ou par paires incluant la lettre latine) faire partie du même jeu de caractères (ce qui est, entre autres, le cas dans le jeu universel de caractères de l'ISO/CEI 10646, dans le jeu de caractères cyrilliques de l'ISO/CEI 8859-5 et dans le jeu de caractères grecs de l'ISO/CEI 8859-7). Lors d'une recherche numérique, il ne serait pas possible de trouver la lettre majuscule latine A dans un texte où elle aurait été camouflée en lettre majuscule grecque ALPHA (A), bien que la forme aurait été repérée visuellement.

non de s'astreindre aux contraintes techniques du passé, qui n'en sont déjà plus dans bien des cas.

Remarques :
Le répertoire minimal de caractères requis au sein de l'Administration gouvernementale est celui de la norme ISO/CEI 8859-15 (Traitement de l'information – Jeux de caractères graphiques codés sur un seul octet : Alphabet latin n° 9 : 1999) de l'Organisation internationale de normalisation (ISO) et de la Commission électrotechnique internationale (CEI).

Dans le jeu de caractères codés de la norme internationale ISO/CEI 8859-15 se trouve l'ensemble des caractères nécessaires à l'écriture intégrale des lettres et des signes de ponctuation essentiels du français et les caractères propres à 27 langues de l'Europe de l'Ouest, en plus du symbole de l'euro (€), bien qu'il faille recourir à un jeu plus complet, comme celui du jeu universel de caractères, pour produire des textes de qualité typographique soignée (comportant, par exemple, des tirets sur cadratin et sur demi-cadratin) ou des textes à graphie savante (par exemple *qāt*, graphie savante du mot *qat*) de même que des textes utilisant d'autres systèmes d'écriture que l'écriture latine.

1. Cette exigence se fait dans le respect optimal des systèmes existants tout en permettant de répondre aux besoins actuels.
2. Pour les besoins du standard sur les adresses de courriel (SGQRI 044), les caractères constituant une adresse de courriel doivent pour l'instant être limités à un sous-ensemble du répertoire de l'IRV de la norme ISO/CEI 646 (Technologies de l'information – Jeu ISO de caractères codés à 7 éléments pour l'échange d'information) de l'ISO et de la CEI. Ce sous-ensemble est constitué des lettres minuscules et majuscules de l'alphabet (de a à z), des chiffres allant de 0 à 9 et du caractère « - » (tiret).

5. Les jeux de caractères codés qui peuvent être utilisés sont les suivants :

Jeux de caractères d'usage général :
Les jeux de caractères définis ci-après comprennent l'ensemble des caractères qui permettent un soutien intégral du français. Les deux jeux de caractères ISO qui suivent sont recommandés pour toutes les données textuelles d'usage général destinées à être transmises au grand public. Les autres codes sont des jeux de caractères privés équivalents : la majeure partie des données informatiques gouvernementales peut difficilement être convertie rapidement, compte tenu, notamment qu'elles sont stockées sur des ordinateurs de grande puissance compatibles à ceux d'IBM.. Le choix parmi les jeux suivants dépendra donc de l'usage envisagé et des contraintes techniques existantes. L'essentiel pour l'interopérabilité reste le soutien du répertoire minimal requis.

a) le jeu de caractères de la norme internationale ISO/CEI 10646 (Technologies de l'information, Jeu universel de caractères codés sur plusieurs octets [JUC]) de l'Organisation internationale de normalisation (ISO) ;

Précisions techniques :
1. Pour les applications générales, c'est le jeu universel de caractères (JUC), qui correspond à la norme internationale ISO/CEI 10646 et au standard du consortium Unicode désigné sous l'appellation The Unicode Standard, qu'il faut préconiser. Le jeu de caractères d'Unicode est identique à celui de la norme internationale ISO/CEI 10646 en ce qui concerne le répertoire, les numéros des caractères et les formes de codage. Le standard Unicode contient toutefois des

spécifications additionnelles qui le rendent nécessaire, notamment les propriétés des caractères et la notion d'équivalence canonique.

Le répertoire du JUC contient beaucoup plus de caractères que ceux qui sont essentiels au français. Son grand répertoire le rend propre à tous les types de communications écrites. La norme internationale ISO/CEI 10646 en garantit la pérennité et la stabilité, le rendant propre à l'archivage de données à long terme.

L'une des propriétés du JUC est de permettre la représentation des lettres accentuées de deux façons :

- en utilisant un caractère dédié, dit précomposé, compris dans le répertoire (par exemple U+00E7 ç) ;

- en utilisant une suite combinatoire formée du caractère de base suivi d'un caractère combinatoire représentant l'accent (par exemple U+0063 + U+0327).

Tout logiciel conforme à Unicode **doit** interpréter ces deux représentations de façon identique, mais pour plusieurs processus (notamment la comparaison de chaînes, base de nombreux processus plus complexes), il importe de n'avoir qu'une représentation dite canonique. L'annexe n° 15 du standard Unicode (Unicode Normalization Forms [UAX#15]) propose quatre formes normalisées propres à remplir le rôle de forme canonique. Ce standard impose l'utilisation de la forme appelée *NFC* dans l'UAX#15. En d'autres mots, cette forme est telle qu'il faut utiliser la représentation précomposée de toute lettre accentuée, en autant qu'elle existe dans le répertoire (toute les lettres accentuées du français y figurent).

Il existe trois formes normatives de codage du JUC, normalisées tant dans le standard Unicode que dans la norme internationale ISO/CEI 10646 : UTF-8 (Universal Character Set Transformation Format 8, en français « format transformé de codage 8 du jeu universel de caractères »), UTF-16 et UTF-32, dont les unités de codage sont respectivement de 8, 16 et 32 bits. Dans chaque cas d'utilisation, il faut choisir une de ces formes ; le présent standard n'impose pas de choix mais propose les critères qui suivent.

- La forme UTF-8 doit être utilisée pour la transmission de données textuelles dans tous les cas où l'environnement technique est non homogène (dans un environnement parfaitement homogène, il n'est pas nécessaire de transcoder les jeux de caractères). Il faut noter en particulier que l'*Internet Engineering Task Force* (IETF) en a fait le codage prioritaire pour tous les protocoles Internet (cf. RFC 2277, SMTP Service Extension for 8bit-MIME transport) et qu'il a fait l'objet d'une normalisation par cet organisme (cf. RFC 2279, UTF-8, A transformation format of ISO 10646). Cette forme est à privilégier pour tous les échanges d'information entre toutes les plates-formes qui ne partagent pas le même jeu de caractères codés.

- La forme UTF-16 est recommandée pour le *stockage* et le *traitement* de données textuelles.

  C'est notamment la forme utilisée par l'environnement Java et les versions récentes des systèmes d'exploitation Windows et MacOS.

- La forme UTF-32 est la seule forme utilisable lorsque des contraintes techniques imposent le choix d'un codage de longueur fixe, c'est-à-dire que chaque caractère est représenté par le même nombre d'octets. UTF-32 utilise quatre octets par caractère, UTF-16 deux ou quatre, et UTF-8 un, deux, trois ou quatre.

2. Pour les ordinateurs de grande puissance de type IBM, il existe une version non normalisée, différente d'UTF-8, appelée EBCDIC-UTF-8 (EBCDIC, Extended Binary-Coded-Decimal Interchange Code). Unicode est un standard qui a pour but d'inclure des considérations pratiques d'implantation du JUC, mais UTF-8 doit être adapté pour utilisation sur ces grands systèmes pour ne pas nuire au fonctionnement de l'infrastructure.

b) le jeu de caractères de la norme internationale ISO/CEI 8859-15 (alphabet latin n° 9) de l'ISO ;

Précisions techniques :
Lorsqu'une contrainte technique incontournable impose l'utilisation d'un jeu de caractères à 8 bits, il faut de préférence utiliser la norme internationale l'ISO/CEI 8859-15, aussi connue sous le nom « alphabet latin n° 9 », ou encore désignée dans Internet comme « ISO 8859-15 » (nom MIME [Multipurpose Internet Mail Extension]). Ce jeu contient tous les caractères essentiels du français et aussi le symbole de l'euro (€) mais il reste limité, de par sa nature de jeu à 8 bits (non-soutien de caractères nécessaires à la typographie soignée, de langues étrangères utilisant différents systèmes d'écriture et de caractères mathématiques et scientifiques).

Il n'y a qu'une seule forme de codage de la norme internationale ISO/CEI 8859-15, dans laquelle chaque caractère est représenté par un octet de valeur égale au numéro du caractère.

Remarques :
1. Lorsque le facteur déterminant est la plus grande compatibilité possible avec le plus grand nombre d'interlocuteurs (par exemple envoi de messages ou diffusion de pages Web au grand public), le jeu de caractères qui est le plus répandu en Occident est la norme internationale ISO/CEI 8859-1, aussi connu sous les noms « alphabet latin n° 1 », ISO Latin-1 ou ISO-8859-1 (nom MIME). Ce jeu de caractères s'est imposé au fil des ans et a été adopté implicitement pour le français et plusieurs autres langues occidentales dans la messagerie Internet, les forums de messageries (Usenet), les pages Web, les systèmes Unix, etc. Son grand défaut est cependant d'être un jeu de caractères à 8 bits d'un répertoire forcément encore plus limité que l'alphabet latin n° 9, et il est également dépourvu de trois caractères essentiels à l'orthographe du français (œ, Œ, Ÿ) et du symbole de l'euro (€). L'apparition du symbole € en Europe fera en sorte que d'autres jeux de caractères, dont l'alphabet latin n° 9, seront de plus en plus fréquemment utilisés dans tous ces contextes.

2. Le codage décrit dans la norme internationale ISO/CEI 8859-15 est identique au codage décrit dans la norme internationale ISO/CEI 8859-1, sauf pour huit caractères. Les huit caractères « ¤ ´ ¦ ¨ ¸ ¼ ½ ¾ » de la norme internationale ISO/CEI 8859-1 ont été remplacés respectivement par « € Ž Š š ž Œ œ Ÿ » (le symbole de l'euro, quatre caractères finnois et trois caractères français) pour rectifier les erreurs historiques en ce qui concerne le soutien intégral du français et du finnois, de même que pour introduire le symbole de l'euro dans les jeux de caractères à 8 bits de la série de normes internationales ISO/CEI 8859.

3. Compte tenu que les anciennes banques de données textuelles codées selon l'alphabet latin n° 1 ne pouvaient pas contenir les nouveaux caractères alphabétiques de l'alphabet latin n° 9 qui sont nécessaires au soutien intégral du français, il n'y a pas lieu de convertir les données dans la plupart des cas, les caractères abandonnés pouvant alors être considérés comme des erreurs de frappe, ce qu'ils sont dans la plupart des cas, compte tenu du peu de signification qu'ils ont. Une simple redéclaration du codage suffit dans la majorité des cas.

Jeux de caractères d'usage restreint
Pour des usages restreints, notamment lorsqu'il ne s'agit pas d'envisager la transmission des données textuelles ou leur stockage permanent ou semi-permanent dans le but d'un accès ultérieur par des ordinateurs autres que celui qui effectue le stockage, les jeux de caractères autres que ceux désignés comme d'usage général précédemment peuvent être utilisés. C'est le cas, par exemple, lorsque des données textuelles sont créées, traitées et stockées de façon temporaire dans un même ordinateur qui emploie un jeu de caractères particulier au fabricant. Si les données doivent être transmises ou stockées de façon permanente, il conviendra de les transcoder en caractères d'usage général.

Ce standard ne précise les jeux de caractères d'usage restreint que pour les ordinateurs de grande puissance de type IBM et pour les ordinateurs (postes de travail individuels et serveurs) utilisant des versions récentes de Windows, compte tenu que la majorité des données informatiques du gouvernement du Québec sont stockées dans des ordinateurs de ce genre. Hormis ces jeux de caractères, nous trouvons aussi des pages de codes propres aux micro-ordinateurs utilisant les systèmes d'exploitation Windows ou MacOS de même que d'autres pages de code EBCDIC sur ordinateurs de grande puissance IBM ou compatibles.

La seule contrainte imposée aux jeux de caractères d'usage restreint est que leur répertoire doit contenir au minimum tous les caractères essentiels au français. L'interopérabilité dans les deux sens d'un échange de données ne sera toutefois garantie que pour les caractères du répertoire de la norme internationale ISO/CEI 8859-15, ce qui est le cas des deux derniers jeux de caractères énoncés dans le présent standard et qui suivent cette mise en contexte.

c) le jeu de caractères de Microsoft désigné sous l'appellation MS-1252 (dans sa version postérieure à 1998), à l'exclusion des 32 caractères additionnels qui ne font pas partie du répertoire de la norme internationale ISO/CEI 8859-15 ;

Précisions techniques :
L'inconvénient de la référence à ce jeu de caractères est que ce dernier a été augmenté par Microsoft au fil des ans sans laisser de trace de la version. À partir de Windows 98, la version définitive de ce jeu de caractères est devenue standard sous Windows ; pour Windows 95, un correctif est disponible. Ce dernier jeu de caractères comprend 223 caractères graphiques alors que la norme internationale ISO/CEI 8859-15 n'en comprend structurellement que 191 (dans la structure standard de toutes les normes internationales ISO de jeux de caractères codés sur 8 bits, les 32 caractères supplémentaires sont réservés aux caractères de commande, ce que Microsoft n'a pas respecté). En conséquence, 32 caractères sont perdus s'ils sont convertis dans un jeu de caractères à 8 bits de structure ISO ou dans un jeu de caractères EBCDIC).

d) le jeu de caractères d'IBM désigné sous l'appellation EBCDIC 924.

Précisions techniques :
Il s'agit de l'équivalent EBCDIC de la norme internationale ISO/CEI 8859-15. Il faut souligner ici qu'avant 1991 le gouvernement du Québec utilisait majoritairement un jeu de la famille des jeux de caractères EBCDIC, de variante dite « française canadienne », qui a été converti sur recommandation d'IBM au jeu de caractères EBCDIC 037, version 1, dite nord-américaine ou « bilingue canadienne ». IBM a plus tard recommandé à tous les pays européens – et au Canada – de passer au jeu EBCDIC 500, qui comprend les mêmes caractères mais sous un codage différent pour certains signes utilisés dans les langages de programmation. Le jeu EBCDIC 924 nécessitera une conversion (mais pas pour les caractères alphabétiques) si certains grands systèmes continuent d'utiliser le jeu de caractères EBCDIC 037 (pour les cas où

le langage de programmation PL/1 continue d'être compilé après modification des programmes, ce qui est relativement rare au gouvernement du Québec). Le jeu EBCDIC 037 comporte intégralement tous les caractères du français et est autrement compatible, pour ce qui est du texte français, aux bases de données textuelles implantées au gouvernement du Québec (voir aussi le standard sur le tri alphabétique et la recherche de chaînes de caractères [SGQRI 004] de l'Administration gouvernementale).

Considérations spéciales pour Internet

1. Nom de domaine Internet et adresse de courriel

Il faut se référer aux standards sur les noms de domaine Internet (SGQRI 021) et sur les adresses de courriel (SGQRI 044) de l'Administration gouvernementale, qui traitent respectivement des noms de domaine Internet et des adresses de courriel. Ces standards précisent certaines considérations (qui sont sujettes à modification) concernant l'utilisation des jeux de caractères.

2. Adresse de site Web

Les adresses de site Web sont des URI, tels que définis par le standard de l'IETF désigné sous l'appellation RFC 2396 (Uniform Resource Identifiers (URI) : Generic Syntax). Ce standard précise qu'un URI ne peut contenir qu'un sous-ensemble des caractères de l'*American National Standards Institute* (ANSI) désigné sous l'appellation ASCII (American Standard Code for Information Interchange, les caractères de numéros inférieurs à 128 dans la norme internationale ISO/CEI 10646). Il est toutefois prévu qu'un URI puisse contenir d'autres caractères, à condition que les octets représentant ces caractères codés soient à leur tour *surcodés* sous la forme %HH, où HH est la notation hexadécimale de la valeur de l'octet. Rien dans le RFC ne précise toutefois le codage des caractères non ASCII, ce qui rend leur utilisation difficile et incertaine.

3. Dans certains contextes, nommément dans les documents HTML (American Standard Code for Information Interchange) et XML (Extensible Markup Language), un mécanisme bien défini est prévu pour le traitement des URI contenant des caractères non ASCII (cf. l'annexe B.2.1 du standard du *World Wide Web Consortium* (W3C) désigné sous l'appellation HTML 4.01 et la section 4.2.2 du standard du W3C désigné sous l'appellation XML 1.0). Dans ces cas, il est précisé que les URI sont d'abord exprimés en UTF-8, puis que les octets non ASCII sont surcodés sous la forme %HH mentionnée ci-dessus. Ce sont les fureteurs qui se chargent de ces opérations avant de transmettre l'URI à un serveur ; les documents HTML et XML peuvent donc contenir des URI avec des caractères autres que ceux de l'ASCII. Pour que le mécanisme fonctionne, il est nécessaire que les serveurs recevant ces URI sachent les interpréter.

4. Serveur de courrier électronique SMTP/MIME (SMTP, Simple Mail Transfer Protocol)

Tout comme dans le cas des adresses de messagerie, le standard de l'IETF désigné sous l'appellation *RFC 2821* (Simple Mail Transfer Protocol) régissant les serveurs de courriel a été conçu il y a fort longtemps essentiellement pour servir les besoins d'utilisateurs américains. Ce standard original imposait donc la restriction du répertoire de caractères utilisable au seul jeu ASCII. Depuis, des extensions ont été prévues permettant le transport d'autres caractères – dont ceux nécessaires au français – mais l'implantation de ces extensions reste facultative dans les standards de l'IETF. Pour assurer le bon fonctionnement de la messagerie en français, ce standard impose donc la conformité à certaines parties optionnelles des standards et l'implantation de certains mécanismes :

- tous les serveurs de transport de courrier électronique doivent être conformes au standard

RFC 2821, y compris le mécanisme d'extension EHLO[2] ;

- parmi les extensions possibles, les serveurs doivent au minimum implanter l'extension 8bit-MIMEtransport (8BITMIME) décrite dans le standard de l'IETF désigné sous l'appellation RFC 1652 (SMTP Service Extension for 8bit-MIMEtransport) ;

- de surcroît, les serveurs doivent disposer d'un mécanisme de transformation d'un message reçu en 8BITMIME en un message 7bit MIME pour transmission à un éventuel serveur externe à l'appareil gouvernemental qui n'implanterait pas l'extension 8BITMIME (cf. section 3 du standard RFC 1652) ;

- finalement, les serveurs devraient transformer les messages textuels reçus en 7bit MIME en messages 8BITMIME avant de les transmettre à des serveurs ou des clients 8BITMIME.

SECTION III : DISPOSITIONS TRANSITOIRES ET FINALES

S.-s. 1 – Mesures transitoires

Mise en contexte :
Les ordinateurs centraux du gouvernement du Québec qui utilisent des caractères accentués ont utilisé de 1991 à 2003 le jeu de caractères d'IBM désigné sous l'appellation EBCDIC 037 CECP, qui permet, après transcodage, un échange sans perte d'information dans les deux sens, avec le jeu de caractères codés de la norme internationale ISO/CEI 8859-1. Ce dernier jeu de caractères codés permettait de soutenir le français avec quelques restrictions mineures. Depuis 1998, il existe un jeu de caractères codés (norme internationale ISO/CEI 8859-15) qui soutient le français intégral et qui a été normalisé pour prévoir le symbole de l'euro (€). Le répertoire du jeu de caractères codés IBM EBCDIC 924 correspond exactement à ce répertoire minimal. Les données alphabétiques du jeu de caractères EBCDIC 037 CECP sont transposables sans conversion vers ce nouveau jeu de caractères, qui ne comporte que certaines différences qui pourraient affecter l'affichage de programmes sources (notamment en programmation PL/1, un langage pratiquement disparu au gouvernement du Québec). Il faudra, par contre, prévoir certains ajustements mineurs de microcode pour l'affichage sur certains types d'équipement (contrôleurs de terminaux, notamment les contrôleurs IBM 3174), ces ajustements ne posant pas a priori de risque grave de dysfonctionnement des applications.

Sous MS-Windows, le répertoire du jeu de caractères codés de la norme internationale ISO/CEI 8859-15 (répertoire minimal) est transposable dans les jeux de caractères MS-1252 et Unicode (selon les versions du système d'exploitation). Ces deux jeux de caractères sont donc conformes dans la mesure où, pour les échanges avec des applications utilisant d'autres jeux de caractères, nous nous assurons que le répertoire saisi ne déborde pas le répertoire du jeu de la norme internationale ISO/CEI 8859-15. Le choix entre le jeu MS-1252 ou le jeu Unicode dans une application de la plate-forme Windows relève de chaque ministère ou organisme.

Sur plate-forme Unix, le jeu de caractères de la norme internationale ISO/CEI 8859-15 est le

[2] Commande évoluée du protocole SMTP qui demande à ce que le serveur destinataire qui a indiqué au préalable qu'il acceptait cette commande renvoie au serveur requérant des données sur la nature des services qu'il peut rendre.

choix à adopter directement en lieu et place de celui de la norme internationale ISO/CEI 8859-1. Les données alphabétiques de ce dernier jeu de caractères sont utilisables sans conversion avec le nouveau jeu de caractères.

Pour ce qui est des tris, il faut se référer au standard sur le tri alphabétique et la recherche de chaînes de caractères (SGQRI 004) de l'Administration gouvernementale. En général, les seuls changements à ce titre n'ont trait qu'à l'identification du jeu de caractères lui-même dans les tables de classement et à la mise à jour de ces tables, pour des caractères peu susceptibles de se trouver déjà à grande échelle dans les fichiers informatiques.

Les nouveaux protocoles d'Internet, quant à eux, nécessitent une implantation du jeu universel de caractères, et ceci signifie une prise en charge automatique du répertoire minimal défini dans ce standard.

6. Tout jeu de caractères codés utilisé avant la date d'entrée en vigueur du présent standard doit être conforme aux exigences de la section II au plus tard deux ans après cette date.

S.-s. 2 – Révision

7. Au plus tard cinq ans après l'entrée en vigueur de ce standard, le ministère des Services gouvernementaux doit, de concert avec les ministères et les organismes, en évaluer la mise en œuvre et conseiller le ministre des Services gouvernementaux quant à l'opportunité d'y apporter des modifications en vue d'une proposition au Conseil du trésor.

S.-s. 3 – Date d'entrée en vigueur

8. Ce standard entre en vigueur le 11 décembre 2006.

Explication :
Tout jeu de caractères utilisé après l'entrée en vigueur de ce standard doit y être conforme.

RENSEIGNEMENTS COMPLÉMENTAIRES

R.C. 1 – Autres sigles et définitions

R.C. 1.1 –Sigles

ASCII
American Standard Code for Information Interchange. Jeu de caractères codés sur 7 bits correspondant à la norme internationale ISO/CEI 646.

CEI
Commission électrotechnique internationale

CECP
Country-Extended Code Page. Série de jeux de caractères IBM étendant le code EBCDIC initial pour assurer la correspondance avec certaines normes de l'ISO relatives aux jeux de caractères codés sur huit bits.

EBCDIC
Extended Binary-Coded-Decimal Interchange Code. Série de jeux de caractères codés sur 8 bits et utilisés dans des environnements compatibles aux ordinateurs de grande puissance IBM.

EHLO
Commande évoluée du protocole SMTP qui demande à ce que le serveur destinataire qui a indiqué au préalable qu'il acceptait cette commande renvoie au serveur requérant des données sur la nature des services qu'il peut rendre.

HTML
HyperText Markup Language. Langage de balisage de documents.

IETF
Internet Engineering Task Force. Communauté volontaire de personnes intéressées au développement de l'architecture Internet et de ses standards (http://www.ietf.org/overview.html).

IRV
International Reference Version. Version de référence internationale de la norme internationale ISO/CEI 646 (correspondant depuis 1991 à l'ASCII, qui est la version nationale américaine de cette norme internationale ; l'ASCII est aussi connu sous l'acronyme de IA5, International Alphabet 5).

ISO
Organisation internationale de normalisation (ISO est un symbole provenant de la racine grecque *isos* – qui signifie *même*) ; en anglais : *International Organization for Standardization*.

JUC
Jeu universel de caractères, décrit dans la norme internationale ISO/CEI 10646 et le standard Unicode (équivalent anglais : UCS).

MIME
Multipurpose Internet Mail Extension

M/O
Ministères et organismes de l'Administration gouvernementale

NFC
Normalization Form C. Forme normalisée C sous Unicode choisissant comme équivalence canonique un caractère précomposé plutôt qu'une séquence combinatoire, quand il existe de telles équivalences.

RFC
Request for Comments. Chaque RFC fait partie d'une série de notes techniques et organisationnelles dans Internet, série qui a débuté en 1969. Un RFC peut devenir un standard Internet ; entre-temps, plusieurs RFC sont appliqués en fonction des besoins des utilisateurs de l'Internet.

SMTP
Simple Mail Transfer Protocol. Protocole utilisé pour les échanges de messagerie dans Internet.

UCS
Universal multiple-octet-coded Character Set. Jeu universel de caractères codés sur plusieurs octets, défini dans la norme internationale ISO/CEI 10646 (équivalent français : JUC).

URI
Uniform Resource Identifier. Identificateur de ressources uniforme, c'est-à-dire une chaîne de caractères identifiant une ressource matérielle ou abstraite ; plus généralement, un URI est l'adresse d'un site Web.

UTF
UCS Transformation Format. Série de formats transformés du jeu universel de caractères codés sur plusieurs octets, dont le but est de transposer ce jeu de caractères de manière compatible avec les environnements techniques existants.

XML
Extensible Markup Language. Langage de balisage de documents défini par le consortium W3C, et plus évolué que le langage HTML.

R.C. 1.2 –Définitions

Équivalence canonique
Équivalence entre un caractère précomposé et une séquence combinatoire correspondante dans le standard Unicode.

Exemple :
Le caractère *LETTRE MINUSCULE LATINE E ACCENT AIGU* est canoniquement équivalent à la séquence de deux caractères composée de la *LETTRE MINUSCULE LATINE E* suivie du *DIACRITIQUE ACCENT AIGU*.

Forme de codage
Spécification, en termes d'unités de stockage telles que l'octet (8 bits) ou le seizet (16 bits), de la manière dont les différentes représentations codées des caractères d'un jeu sont stockées dans la mémoire d'un ordinateur.

> **Remarque :**
> Les formes de codage (souvent appelées simplement codages) vont du très simple (par exemple un octet de valeur égale au numéro du caractère) au plus complexe (suite d'octets ou de seizets de longueur variable dépendant de façon non triviale du numéro du caractère).

Version de référence internationale
Jeu de caractères qui sert d'étalon mondial dans la norme internationale ISO/CEI 646.

R.C. 2 – Références bibliographiques

R.C. 2.1 – Références normatives

GOUVERNEMENT DU QUÉBEC. *SGQRI 004, Tri alphabétique et la recherche de chaînes de caractères.* 2006.

GOUVERNEMENT DU QUÉBEC. *SGQRI 021, Noms de domaine Internet.* 2006.

GOUVERNEMENT DU QUÉBEC. *SGQRI 044, Adresses de courriel.* 2006.

INTERNET ENGINEERING TASK FORCE. *RFC 1652*, *SMTP Service Extension for 8bit-MIMEtransport*, http://www.ietf.org/rfc/rfc1652.txt .

INTERNET ENGINEERING TASK FORCE. *RFC 2277*, *SMTP Service Extension for 8bit-MIMEtransport*, http://www.ietf.org/rfc/rfc2277.txt .

INTERNET ENGINEERING TASK FORCE. *RFC 2279*, *UTF-8, A transformation format of ISO 10646*, http://www.ietf.org/rfc/rfc2279.txt .

INTERNET ENGINEERING TASK FORCE. *RFC 2396*, *Uniform Resource Identifiers (URI) : Generic Syntax*, http://www.ietf.org/rfc/rfc2396.txt .

INTERNET ENGINEERING TASK FORCE. *RFC 2821*, *Simple Mail Transfer Protocol*, http://www.ietf.org/rfc/rfc2821.txt .

INTERNET ENGINEERING TASK FORCE. *RFC 2822*, *Internet message format*, http://www.ietf.org/rfc/rfc2822.txt .

INTERNET ENGINEERING TASK FORCE. *RFC 2822*, *Internet Message Format*, http://www.ietf.org/rfc/rfc2822.txt .

ORGANISATION INTERNATIONALE DE NORMALISATION. *ISO/CEI 646 Technologies de l'information – Jeu ISO de caractères codés à 7 éléments pour l'échange d'information*.

ORGANISATION INTERNATIONALE DE NORMALISATION. *ISO/CEI 8859 : 1998, Traitement*

de l'information – Jeux de caractères graphiques codés sur un seul octet, Partie 1 : Alphabet latin n° 1.

ORGANISATION INTERNATIONALE DE NORMALISATION. *ISO/CEI 8859 : 1999, Traitement de l'information – Jeux de caractères graphiques codés sur un seul octet, Partie 15 : Alphabet latin n° 9.*

ORGANISATION INTERNATIONALE DE NORMALISATION. *ISO/CEI 10646 : 2003, Technologies de l'information, Jeu universel de caractères codés sur plusieurs octets (JUC).*

UNICODE CONSORTIUM. *The Unicode Standard, Version 4.1.0*. http://www.unicode.org/versions/Unicode4.1.0/ .

UNICODE CONSORTIUM. UAX#15 *Unicode Normalization Forms, Unicode Standard Annex #15*. http://www.unicode.org/reports/tr15 .

WORLD WIDE WEB CONSORTIUM. *HTML 4.01 Spécification*, décembre 1999, http://www.w3.org/TR/html401/
(Traduction française disponible à l'adresse http://www.la-grange.net/w3c/html4.01/cover.html).

WORLD WIDE WEB CONSORTIUM. *Extensible Markup Language (XML) 1.0.* http://www.w3.org/TR/REC-xml .

R.C. 2.2 Autres références

IBM, *National Language Design Guide Volume 2, National Language Support Reference Manual,* annexe I.

IBM, *Unicode support in EBCDIC based systems*, 1998-08-07.

R.C. 3 – Dérogation aux autres standards du gouvernement du Québec

Sans objet.

R.C. 4 – Conformité au concept d'adaptabilité culturelle et linguistique

Ce standard permet l'adaptation des technologies de l'information à la langue officielle du Québec. Il met aussi en place des moyens qui permettent éventuellement l'adaptation à d'autres langues que le français pour des échanges élargis.

R.C. 5 – Composition du groupe de travail responsable de l'élaboration du standard

Depuis le 18 février 2005, le ministre des Services gouvernementaux assume, en matière de gestion des ressources informationnelles, la responsabilité d'élaborer et de proposer notamment des standards au Conseil du trésor. Au moment des travaux du groupe de travail interministériel, de 2002 à 2004, les personnes suivantes représentaient les ministères et les organismes suivants :

Rédacteur et chargé de projet :
LA BONTÉ, Alain Secrétariat du Conseil du trésor

Membres du groupe :	
ASSAFIRI, Abdallah	Secrétariat du Conseil du trésor
AUDET, Hélène	Société de l'assurance automobile du Québec
BÉLANGER, Jean	Ministère de l'Éducation du Québec
BOURASSA, Guy	Régie des rentes du Québec
BRISSETTE, Normand	Ministère de l'Industrie et du Commerce
CHABOT, Celyn	Ministère de l'Emploi, de la Solidarité sociale et de la Famille
DUMONT, Mario	Ministère de l'Emploi, de la Solidarité sociale et de la Famille
DUSSAULT, Marcel	Ministère de l'Éducation du Québec
FORTIN, Steven	Ministère de l'Emploi, de la Solidarité sociale et de la Famille
GRANDMAISON, Jean	Financière agricole du Québec
HUDON, Yves	Secrétariat du Conseil du trésor
LABERGE, Marc	Secrétariat du Conseil du trésor
LÉGARÉ, Bruno	Ministère des Relations internationales
LOISELLE, Denis	Régie des rentes du Québec
MANDJEE, Azim	Office québécois de la langue française
MICHAUD, Florent	Société immobilière du Québec
MONTMINY, Jacques	Ministère de la Culture et des Communications
POTVIN, Ginet	Secrétariat du Conseil du trésor
ROY, Jean-Jacques	Secrétariat du Conseil du trésor