



Standard du gouvernement du Québec sur les ressources informationnelles

**Cadre commun d'interopérabilité
Volet 4 – Soutien de la langue française et de
l'internationalisation des documents
Version 1.3**

31 mars 2005

Québec 

Avertissement

Ce document ne constitue pas pour l'instant un standard officiel du gouvernement du Québec. Il a franchi l'étape de l'enquête élargie qui précède son approbation par le Conseil du trésor. Son contenu est encore sujet à modification sans préavis.

Toute référence à ce document doit donc obligatoirement inclure la mention « ébauche de standard en cours d'officialisation », en plus du numéro et du nom du standard, ainsi que du numéro et de la date de la version.

Pour obtenir des renseignements supplémentaires à propos de ce document, veuillez vous adresser au Sous-ministériat à l'encadrement des ressources informationnelles (SMERI) du Ministère des Services gouvernementaux au (418) 528-6148.

.

Table des matières

SIGLES ET ACRONYMES	5
AVANT-PROPOS	6
1. SOUTIEN DE L'INTERNATIONALISATION DES DONNÉES ÉCRITES.....	7
1.1 JEU DE CARACTÈRES UNIVERSEL	7
1.2 ENSEMBLE MINIMAL DE CARACTÈRES QUE LE GOUVERNEMENT DU QUÉBEC DOIT SOUTENIR	9
1.3 JEUX DE CARACTÈRES UTILISÉS DANS LES DOCUMENTS XML.....	11
1.4 CAS PARTICULIER DES DOCUMENTS TEXTUELS TECHNIQUES, SCIENTIFIQUES ET MUSICAUX	13
1.5 CAS PARTICULIER DES UNITÉS DE MESURE	14
2. SOUTIEN DE L'INTERNATIONALISATION EN ENTRÉE ET EN SORTIE DES SYSTÈMES INFORMATIQUES	16
2.1 LE CLAVIER	16
2.2 IMPRESSION ET AFFICHAGE	17
2.3 TRAITEMENTS PAR LES SYSTÈMES INFORMATIQUES	18
2.3.1. <i>Saisie</i>	18
2.3.2. <i>Tri, classement et recherche</i>	19
2.3.3. <i>Adaptabilité culturelle et linguistique dans les applications</i>	20
3. SOUTIEN DE L'INTERNATIONALISATION POUR LES DONNÉES VOCALES.....	22
3.1 IDENTIFICATION DE LA LANGUE UTILISÉE DANS UN DOCUMENT DESTINÉ À LA RESTITUTION VOCALE	22
3.2 RECONNAISSANCE DE LA PAROLE	23
3.3 SYNTHÈSE DE LA PAROLE À PARTIR DE TEXTE (SYSTÈMES <i>TEXT-TO-SPEECH</i>).....	24
ANNEXE : DÉFINITIONS.....	26

Table des illustrations

<i>Figure 1 : Ensemble minimal de caractères que les ministères et organismes doivent soutenir.....</i>	<i>10</i>
---	-----------

Sigles et acronymes

API	Application Programming Interface (interface de programmation)
ASCII	American Standard Code for Information Interchange (code ASCII)
CAN/CSA	Canadian Standards Association (Association canadienne de normalisation)
CEI	Commission électronique internationale
CNRS/LIMSI	Centre national de la recherche scientifique/Laboratoire d'informatique pour la mécanique et la science de l'ingénieur
EBCDIC	Extended Binary Coded Decimal Interchange Code (code EBCDIC)
HTML	HyperText Markup Language (langage HTML)
ISO	International Organization for Standardization (Organisation internationale de normalisation)
MathML	Mathematical Markup Language
M/O	Ministères et organismes
SGQRI	Standard du gouvernement du Québec pour les ressources informationnelles
SQCA	Service québécois de changement d'adresse
SSML	Speech Synthesis Markup Language
UTF	Unicode Transformation Format
W3C	Consortium W3C
XML	eXtensible Markup Language (langage XML)

Avant-propos

Le soutien de l'usage de la langue française et de l'internationalisation au sein du gouvernement du Québec, et en premier lieu celui de l'internationalisation de la langue française, passe par un effort de normalisation des M/O (*ministères et organismes*) portant sur chacun des composants décrits dans ce volet.

Ce soutien s'exerce d'abord au regard des *données écrites*, qui, lorsqu'elles sont échangées dans un contexte multilingue, devraient toutes être conformes à la norme du *jeu de caractères universel* (norme internationale ISO/CEI 10646, ou standard Unicode). Mais la pénétration de ce standard dans les systèmes informatiques est liée à sa prise en compte progressive par le marché. Pendant la transition, les données de l'Administration québécoise doivent pouvoir s'appuyer sur *un ensemble minimal comprenant tous les caractères du français intégral*, conformément aux recommandations de l'Office québécois de la langue française. Cet ensemble comprend notamment les accents et les ligatures (ex. : œ).

Ensuite, l'approche porte sur les *entrées et sorties des systèmes informatiques*, puisqu'ils sont les outils privilégiés de création et d'édition des documents. Ces phases du cycle de vie d'un document nécessitent un *clavier*, un *affichage* à l'écran et au besoin des *impressions* papier, tandis que tout traitement informatisé de textes recourt au *tri*, au *classement* et à la *recherche* de caractères.

Enfin, le soutien de l'internationalisation nécessite de normaliser les composants relatifs au traitement vocal, dont les deux étapes en interface avec l'utilisateur sont la *reconnaissance de la parole* et la *synthèse vocale*.

1. Soutien de l'internationalisation des données écrites

1.1 Jeu de caractères universel

Contexte

Avant l'invention d'Unicode, plusieurs systèmes de codage de caractères ont été créés. Les plus répandus sont le code ASCII (*American Standard Code for Information Interchange*) et le code EBCDIC (*Extended Binary Coded Decimal Interchange Code*). Aucun de ces systèmes de codage ne couvre cependant toutes les lettres de la langue française, les signes de ponctuation et les symboles techniques en usage courant.

Ces systèmes de codage sont souvent incompatibles entre eux. Ainsi, deux systèmes peuvent utiliser les mêmes séquences hexadécimales (le même codage) pour deux caractères différents ou utiliser différentes séquences (différents codages) pour le même caractère. Les ordinateurs, et plus particulièrement les serveurs, doivent supporter plusieurs systèmes de codage de caractères, ce qui crée un risque de corruption des données à chaque transfert.

La norme internationale [ISO/CEI 10646](#) (compatible avec le standard [Unicode](#)) est un système universel de codage de caractères. Elle définit une manière cohérente de coder des textes multilingues et facilite l'échange de données textuelles à l'échelle planétaire. En tant que codage normalisé, cette norme internationale constitue un socle ferme pour Internet. Elle est la base de tout logiciel qui doit fonctionner partout. Ce jeu de caractères universel est un sur-ensemble de tous les jeux de caractères utilisés à l'échelle planétaire. Il est donc suffisamment étendu pour être utilisé dans les échanges de textes, y compris les textes inclus dans les formats d'image et les sous-titres formés de caractères dans certains formats vidéo. Cependant, il ne permet pas de coder les images, ni les bandes sonores, ni les vidéos numérisées. Les variantes de polices (taille, format et orientation des caractères) sont également exclues du codage.

Le standard [Unicode](#) est compatible avec la norme internationale [ISO/CEI 10646](#) (jeu universel de caractères codés sur plusieurs octets). La norme ISO/CEI 10646 attribue à chaque caractère un nom (en français et en anglais, nom qui permet la conciliation avec d'autres jeux de caractères) et une valeur de code. Quant au standard Unicode, il attribue aux caractères les mêmes noms (mais uniquement en anglais, les noms français n'en font pas partie) et valeurs de code. Ce jeu de caractères est encore en évolution ; la dernière version, [Unicode 3.2](#), considère certains caractères faisant partie des spécifications de MathML (*Mathematical Markup Language*) et ne figurant pas dans les versions précédentes.

Les données du jeu universel de caractères peuvent être codées sous trois formes : UTF-8, UTF-16 et UTF-32, sur respectivement 8, 16 et 32 bits. Le choix parmi ces formes dépend de l'usage envisagé et des contraintes techniques existantes*.

* Pour plus d'information, consulter le standard SGQRI 3 – *Codage des jeux de caractères*.

Au sein des M/O, certains systèmes informatiques patrimoniaux ne soutiennent pas encore le standard Unicode. Dans ce cas, il convient de s'assurer que l'ensemble minimal de caractères décrit dans la section suivante (correspondant au répertoire de l'ISO/CEI 8859-15) est pris en compte intégralement. Ce jeu de caractères contient tous les caractères essentiels au français.

Orientations

Les M/O doivent se référer au standard SGQRI 3 et favoriser autant que possible l'utilisation du jeu universel de caractères, dans le but :

- d'assurer la stabilité de leurs données en évitant la prolifération des jeux de caractères ;
- d'augmenter l'interopérabilité et l'échange de données au sein du gouvernement du Québec.

À défaut du soutien du jeu universel de caractères, notamment par les systèmes patrimoniaux, l'utilisation d'un ensemble de caractères restreint, comme celui qui est illustré ci-dessous dans la figure 1, est à retenir.

No	Composant	Norme ou standard	Orientation	Références	Commentaires
s10	Jeu de caractères à utiliser au gouvernement du Québec	Unicode ISO/CEI 10646	Retenir	http://www.unicode.org/standard/versions/enumeratedversions.html#Latest http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSN=29819	• Il faut se reporter au standard SGRQI 3. • La norme ISO/CEI 10646 et Unicode sont totalement compatibles. Toutes les deux utilisent un seul code de caractère universel. La norme ISO précise des noms pour chacun des caractères en français et en anglais, alors que le standard Unicode est unilingue anglais. Il spécifie donc uniquement les noms en anglais. Ces noms permettent la conciliation des caractères avec ceux d'autres codages, et leur cohérence est donc essentielle. • Au sein des M/O, certains systèmes informatiques patrimoniaux ne soutiennent pas encore le standard Unicode. Dans ce cas, il convient de s'assurer que l'ensemble minimal de caractères décrit dans la section suivante (correspondant au répertoire de l'ISO/CEI 8859-15) est pris en compte. Ce jeu de caractères contient tous les caractères essentiels au français.

1.2 Ensemble minimal de caractères que le gouvernement du Québec doit soutenir

Contexte

Au sein du gouvernement du Québec, l'ensemble des fonctionnaires travaille en français. Les caractères essentiels qu'ils utilisent dans l'exercice de leurs fonctions sont donc les caractères de la langue française.

L'ensemble minimal de caractères (le plus petit commun dénominateur) répondant aux besoins du gouvernement du Québec est celui figurant dans la norme ISO 8859-15 (ISO-Latin n° 9) et qui est illustré ci-dessous. En effet, tous les caractères de ce code sont transposables dans les différents jeux de caractères prescrits dorénavant pour les M/O du Québec et, si l'on se limite à cet ensemble minimal, le transcodage entre toutes les plates-formes est garanti. C'est donc un gage d'interopérabilité entre les M/O.

Si un document ou une donnée contient des caractères ne figurant pas dans cet ensemble (par exemple des caractères arabes, hébreux, etc.), aucune garantie de transcodage ne pourra être donnée tant que l'Unicode n'est pas implanté partout. Par exemple, si ce document est destiné à être traité ou stocké et que l'application ou l'API (*Application Programming Interface*) de la base de données ne prend pas en compte un jeu de caractères représentant l'arabe, le document sera stocké dans le jeu de caractères de cette base de données, et aucune transformation ne permettra de restituer le document d'origine. Par contre, il est vraisemblable que ce document sera affichable sur bon nombre de postes informatiques, car les systèmes d'exploitation utilisent de plus en plus des jeux de caractères beaucoup plus riches.

SP	0	@	P	`	p			hbr	*	À	Ð	à	ð
!	1	A	Q	a	q			ı	±	Á	Ñ	á	ñ
"	2	B	R	b	r			ç	²	Â	Ò	â	ò
#	3	C	S	c	s			£	³	Ã	Ó	ã	ó
\$	4	D	T	d	t			€	Ž	Ä	Ô	ä	ô
%	5	E	U	e	u			¥	µ	Å	Ö	å	ö
€	6	F	V	f	v			Š	¶	Æ	Ö	æ	ö
'	7	G	W	g	w			Š	·	Ç	×	ç	÷
(	8	H	X	h	x			š	ž	È	Ø	è	ø
)	9	I	Y	i	y			©	¹	É	Û	é	ù
*	:	J	Z	j	z			ª	º	Ê	Ú	ê	ú
+	;	K	[	k	{			«	»	Ë	Û	ë	û
,	<	L	\	l				¬	œ	Ï	Ü	ï	ü
-	=	M	]	m	}			shy	œ	Í	Ý	í	ý
.	>	N	^	n	~			@	Ÿ	Î	Þ	î	þ
/	?	O	_	o				-	¿	İ	ß	ï	ÿ

Figure 1 : Ensemble minimal de caractères que les ministères et organismes doivent soutenir
(description du composant s20)

Orientations

Tous les caractères de ce code sont transposables dans les différents jeux de caractères prescrits dorénavant pour les M/O. C'est donc un gage d'interopérabilité. Par conséquent, les M/O ne doivent pas utiliser de caractères en dehors de cet ensemble s'ils ne veulent pas risquer de perdre de l'information au cours des différents transcodages.

Si un document ou une donnée contient des caractères ne figurant pas dans cette table (par exemple des caractères arabes, hébreux, etc.), aucune garantie ne pourra être donnée dans l'état actuel des choses quant à leur intégrité ou à leur restitution.

No.	Composant	Norme ou standard	Orientation	Références	Commentaires
s20	Ensemble minimal de caractères à soutenir au gouvernement du Québec	SGQRI 3	Retenir	SGQRI 3 Communiquer avec Alain LaBonté, Alain.labonte@msg.gouv.qc.ca	• Ce standard prescrit les codages permettant l'interopérabilité de l'ensemble des caractères décrits dans l'ISO-Latin n° 9, ou ISO/CEI 8859-15. • Ce jeu de caractères permet d'écrire, outre en français, dans les langues suivantes : albanais, allemand, anglais, basque, breton, catalan, danois, espagnol, estonien, féroïen, finnois, frison, gaélique écossais, gaélique irlandais (nouvelle orthographe), galicien, gallois, groenlandais, islandais, italien, latin, luxembourgeois, néerlandais, norvégien, portugais, romanche et suédois. • L'ensemble minimal prescrit est défini de manière indépendante du codage. • Les équipes informatiques doivent s'assurer que la représentation de cet ensemble est prise en compte dans le ou les jeux de caractères qu'elles utilisent.

1.3 Jeux de caractères utilisés dans les documents XML

Contexte

Unicode XML propose le jeu universel de caractères. Un tel outil est extrêmement important, notamment sur Internet, où l'on trouve des textes dans toutes les langues.

Le W3C a formulé un ensemble de recommandations visant des restrictions de l'usage d'Unicode dans un document XML. Pour plus de précisions, les M/O peuvent se référer aux spécifications décrites dans le rapport technique n° 20 produit par le consortium Unicode, à savoir : [Unicode in XML and other Markup Language](#).

Pour éviter les problèmes de reconnaissance et de collision qui risquent de se poser, XML définit certains codes, comme `<`, qui correspond au caractère « < », ou `'`, qui correspond à l'apostrophe.

À titre d'exemple, pour éviter une collision avec la marque d'ouverture d'un élément, on pourra écrire dans un document XML le symbole de *inférieur à* avec le code suivant :

`<p> Ceci est un texte XML valide, car
« inférieur à » a été écrit avec le symbole <.</p>`

Lors de l'analyse du document XML (*parsing*), le code est remplacé par sa valeur et le texte devient :

*Ceci est un texte XML valide, car
« inférieur à » a été écrit avec le symbole <.*

On peut aussi faire référence à tout caractère Unicode avec un code numérique de la forme `&#xxx;` où *xxx* est la valeur de ce caractère dans le standard Unicode (en décimal ou en hexadécimal si cette valeur commence par un *x*). Cette façon de faire, de même que l'usage de tout code symbolique (du type *é* pour représenter un *é*, par exemple), doit par contre être évitée pour les caractères du français intégral, pour ne pas alourdir indûment le poids de stockage des fichiers et pour éviter d'autres ennuis de lisibilité*.

Orientation

Si, grâce à Unicode XML, les caractères échangés sont non ambigus, cela ne veut pas dire que l'affichage de ces caractères est toujours possible, d'autant plus que la totalité des plates-formes amenées à traiter du XML dans la fonction publique québécoise n'offre pas un soutien complet de la technologie Unicode. Les M/O doivent se référer au standard SGQRI 011.

* Voir le standard SGQRI 011 – *Critères minimaux pour la conception d'un site multilingue*, qui traite de ce point.

No.	Composant	Norme ou standard	Orientation	Références	Commentaires
s30	Jeu de caractères utilisable dans un document XML au gouvernement du Québec	Unicode XML	Prendre en compte	Note du W3C http://www.w3.org/TR/unicode-xml/	- Le rapport Unicode in XML and other Markup Languages est une coproduction du consortium Unicode. Sa liste de révisions indique que cette version a été modifiée pour passer d'Unicode 3.1 à 3.2. - Unicode évite les collisions entre les balises, les caractères de contrôle et le style. - Bien qu'Unicode XML permette d'échanger les données de manière non ambiguë, il n'est pas toujours possible d'afficher ses caractères, d'autant plus que la totalité des plates-formes amenées à traiter du XML dans la fonction publique québécoise n'offre pas un soutien complet de la technologie Unicode.
		SGQRI 011	Prendre en compte	SGQRI 003 Communiquer avec Alain LaBonté, Alain.labonte@msg.gouv.qc.ca	- Il faut se référer au standard SGQRI 011. - Il faut éviter de faire référence à tous les caractères avec un nom de code symbolique commençant par « & », afin de ne pas alourdir indûment le poids de stockage des fichiers et d'éviter d'autres ennuis de lisibilité.

1.4 Cas particulier des documents textuels techniques, scientifiques et musicaux

Contexte

Les documents textuels (pages Web incluses) techniques, scientifiques et musicaux utilisent des notations spécifiques (éléments chimiques, formules mathématiques, notation musicale, etc.). Ces notations ne font pas partie de l'ensemble minimal de caractères à utiliser au sein du gouvernement du Québec, et risquent d'être perdues lors des transcodages, si le jeu de caractères à l'arrivée (après transcodage) ne sait pas les représenter.

Pour garantir la restitution des notations spécifiques quel que soit l'environnement, il est recommandé aux M/O de les manipuler sous forme d'images (ce qui en restreint l'édition).

Cependant, si les M/O ont besoin de manipuler ces notations (ex. : équations mathématiques dans un rapport), ils devraient avoir recours au jeu de caractères Unicode 3.2, qui permet de coder un grand nombre de symboles et de les manipuler au travers des langages spécifiques comme MathML 2.0 du W3C (voir <http://www.w3.org/Math/>) et

<<http://www.yoyodesign.org/doc/w3c/mathml2/byalpha.html>>).

Si une équipe informatique décide d'y avoir recours, elle doit toutefois être consciente que le jeu complet n'est pas interopérable dans tous les cas. Ainsi, les caractères ne faisant pas partie de l'ensemble minimal de caractères à soutenir au sein du gouvernement du Québec risquent d'être perdus. À titre exemple, lors du stockage dans une base de données, certains caractères risquent de ne pas être affichés dans les versions anciennes de fureteurs, à moins que ces dernières ne supportent l'Unicode.

Orientations

Dans les documents textuels, les M/O peuvent manipuler les notations spécifiques sous forme d'images.

Si un document contient de telles notations qui doivent être manipulées ou éditées, les M/O peuvent avoir recours aux langages spéciaux, comme MathML 2.0, en ayant conscience que ces langages recourent au jeu de caractères défini par Unicode 3.2, ce qui limite l'interopérabilité aux plates-formes soutenant Unicode.

No.	Composant	Norme ou standard	Orientation	Références	Commentaires
s40	Format des notations spécifiques pour les documents scientifiques, techniques et musicaux	Unicode 3.2	Retenir	http://www.unicode.org/standard/versions/enumeratedversions.html#Latest	Unicode 3.2 comprend un grand nombre de symboles mathématiques pouvant être manipulés par le langage MathML 2.0 du W3C.
		Voir composant e220	Retenir	Volet 2 du cadre commun d'interopérabilité	Pour garantir la restitution des notations spécifiques (par exemple : équations mathématiques) quel que soit l'environnement, il est recommandé aux M/O de les manipuler sous forme d'images (ce qui en restreint l'édition).

1.5 Cas particulier des unités de mesure

Contexte

Étant donné que le gouvernement du Québec a adopté le système international des unités de mesure, la saisie et la représentation des éléments numériques (date, heure, poids et mesures) doivent donc suivre l'usage international. Tous les symboles correspondant aux unités de mesure existent dans l'ensemble minimal de caractères que les M/O doivent utiliser et qui est illustré dans la figure 1.

Orientation

Le gouvernement du Québec a adopté le système international (système métrique) en 1970 ; par conséquent, les M/O doivent respecter l'usage international en ce qui concerne les unités de mesure.

No.	Composant	Norme ou standard	Orientation	Références	Commentaires
s50	Affichage des unités de mesure	Ensemble minimal de caractères à soutenir au sein des M/O (composant s20)	Retenir	Voir composant s20	Les unités du système d'unités international sont affichables en n'ayant recours qu'à l'ensemble minimal de caractères du composant s20.

2. Soutien de l'internationalisation en entrée et en sortie des systèmes informatiques

Les unités d'entrée et de sortie d'ordinateurs doivent permettre l'affichage, la saisie et la reproduction de tous les caractères du français correspondant au composant s20. Le cas échéant, la saisie et le traitement de données vocales doivent permettre la reconnaissance et la reproduction du français parlé au Québec, en France et en règle générale dans toute la Francophonie.

2.1 Le clavier

Contexte

La saisie à partir d'un clavier doit non seulement permettre à un utilisateur d'utiliser tous les caractères de la langue française ainsi que les caractères latins de l'Europe de l'Ouest, mais également de changer de poste de travail sans avoir à apprendre à utiliser un nouveau clavier. La partie alphanumérique de tous les claviers doit comporter 48 touches ou plus et doit être conforme au moins au niveau B de la norme CAN/CSA Z243.200-92.

Orientation

Les ministères et organismes doivent se référer au standard SGQRI 001 relatif au clavier québécois.

No.	Composant	Norme ou standard	Orientation	Références	Commentaires
s60	Clavier	SGQRI 001 CAN/CSA Z243.200-92	Retenir	SGQRI 001 Communiquer avec Alain LaBonté, Alain.labonte@msg.gouv.qc.ca http://www.scc.ca/	• Il faut tenir compte du standard SGQRI 001. • La partie alphanumérique de tous les claviers doit comporter 48 touches ou plus et doit être conforme au moins au niveau B de la norme CAN/CSA Z243.200-92.

2.2 Impression et affichage

Contexte

Il existe deux manières de représenter des caractères à l'écran et dans les imprimantes : bitmap (tableau de points) et vectorielle. La représentation bitmap consiste à placer chaque lettre dans un rectangle et à donner les points de ce rectangle à encre par opposition à ceux qui restent blancs. La représentation vectorielle est pour sa part plus abstraite et élaborée, et utilise des fonctions mathématiques géométriques (traits et courbes) pour décrire les caractères.

Les polices bitmap sont très précises pour les petites tailles, alors que les polices vectorielles sont indispensables pour un vrai travail de qualité : elles permettent de faire varier la taille, la résolution ou la forme des caractères sans perte de qualité.

Les polices vectorielles conçues selon la norme True Type sont principalement utilisées pour l'affichage sur écran dans des systèmes implantés largement sur le marché.

Dans le cadre de l'internationalisation, les fonctions d'impression conformes au standard Unicode sont nécessaires pour pouvoir imprimer les caractères non latins (tels que le cyrillique, l'arabe, l'hébreu, etc.). Cependant, elles ne sont actuellement pas disponibles sur l'ensemble des imprimantes de la fonction publique québécoise.

L'interopérabilité des formats de papier peut poser des problèmes, notamment quand un document est édité dans un format non pris en compte par l'unité d'impression (ex. : format A4 sur une imprimante configurée pour le papier « Lettre US »).

Ce point est traité dans le standard SGQRI 047 – *Paramètres minimaux de localisation des logiciels*.

Orientations

Afin de pouvoir afficher et imprimer tous les caractères du français intégral à l'aide des polices appropriées, les M/O doivent s'assurer de l'intégrité de l'ensemble minimal de caractères illustré dans la figure 1.

Pour les acquisitions futures, les M/O doivent privilégier des produits prenant en compte le standard Unicode. Ce dernier s'impose pour l'affichage des documents multilingues ou des symboles techniques. Il est également à noter que, pour des raisons de qualité, les polices vectorielles sont à préférer aux polices bitmap.

L'interopérabilité des formats de papier peut également poser des problèmes lorsqu'un document est dans un format non pris en compte par l'unité d'impression. Le standard SGQRI 047 traite de ce problème.

No.	Composant	Norme ou standard	Orientation	Références	Commentaires
s70	Impression et affichage d'un document	Unicode	Retenir	http://www.unicode.org/standard/versions/enumeratedversions.html#Latest	• Pour les acquisitions futures, les M/O doivent privilégier des produits prenant en compte le standard Unicode. • Unicode est le standard qui s'impose pour l'affichage des documents multilingues ou des symboles techniques.
		ISO 8859-15	Retenir	ISO/CEI-885915	• Afin d'afficher et d'imprimer toutes les lettres de la langue française, il faut s'assurer de l'intégrité de l'ensemble de caractères figurant dans la norme ISO/CEI 8859-15 (ISO-Latin n° 9).
s90	Police d'affichage	True Type	Retenir	http://www.apple.com/	• True Type assure une meilleure qualité de représentation pour les écrans. • C'est un format propriétaire, mais très répandu.
s100	Format de papier	SGQRI 047	Éligible dans une version ultérieure	SGQRI 047 Communiquer avec Alain LaBonté, Alain.labonte@msg.gouv.qc.ca	• L'interopérabilité des formats de papier peut également poser des problèmes lorsqu'un document est dans un format non pris en compte par l'unité d'impression. Le standard SGQRI 047, en cours d'élaboration, devrait régler ce problème. • Pour plus de précisions, communiquer avec Alain LaBonté à l'adresse suivante : <alb@sct1.gouv.qc.ca>.

2.3 Traitements par les systèmes informatiques

2.3.1 Saisie

Contexte

La saisie de l'information textuelle doit se faire (notamment sur le Web) en utilisant toute l'étendue des caractères du français intégral, et les données ainsi saisies doivent pouvoir être préservées dans leur plus stricte intégrité (aucune élimination des accents ou transformation de ligatures en caractères multiples).

Orientation

Les M/O doivent s'assurer que tous les caractères du français intégral* peuvent être saisis.

No.	Composant	Norme ou standard	Orientation	Références	Commentaires
s110	Saisie	Ensemble minimal de caractères à soutenir au sein des M/O (composant s20)	Retenir	Voir composant s20	Les M/O doivent pouvoir saisir tous les caractères du français intégral, à savoir tous les caractères faisant partie de l'ensemble minimal de caractères à soutenir au sein du gouvernement du Québec.

2.3.2. Tri, classement et recherche**Contexte**

Les données textuelles nécessitent souvent d'être triées en ordre alphabétique ou numérique afin d'être traitées selon un ordre établi par l'utilisateur ou l'application. Pour cela, il est impératif que la fonction de tri tienne compte de la casse (majuscules et minuscules), des accents et des caractères spéciaux, et ce, conformément à la norme internationale ISO/CEI 14651:2001.

Le classement alphanumérique fait par une application dans un but d'échange avec d'autres applications ou de présentation doit se faire en conformité avec les attentes des utilisateurs francophones du Québec, selon les critères établis pour l'ordre alphabétique des dictionnaires français, critères parfaitement décrits et pris en compte dans la norme canadienne CAN/CSA Z243.4.1, laquelle constitue un *delta* conforme à la norme internationale ISO/CEI 14651.

La recherche d'information textuelle (notamment par mots clés, sur le Web) devra permettre les requêtes imprécises sans accents (pour retrouver les mots accentués) et les requêtes précises, avec accents, selon les principes établis dans la norme internationale de classement pour la comparaison de chaînes de caractères (ISO/CEI 14651).

Pour plus de précisions sur la méthode de tri, de classement et de recherche, les M/O doivent se référer au standard SGQRI 004.

* Voir la figure 1.

Orientations

Les M/O doivent s'assurer que la fonction de tri tient compte de la casse, des accents et des caractères spéciaux et que le classement se fait en conformité avec les attentes des utilisateurs francophones du Québec, selon les critères établis pour l'ordre alphabétique des dictionnaires français. Quant à la recherche, elle devra se faire par mots clés, sans égard à la casse ni à l'accentuation.

No.	Composant	Norme ou standard	Orientations	Références	Commentaires
s120	Tri, classement et recherche	SGQRI 004	Prendre en compte	SGQRI 004 Communiquer avec Alain LaBonté, Alain.labonte@msg.gouv.qc.ca	• Il faut se référer au standard SGQRI 004. • Le tri doit tenir compte de la casse, des accents et des caractères spéciaux. • Le classement se fait en conformité avec les attentes des utilisateurs francophones du Québec, selon les critères établis pour l'ordre alphabétique des dictionnaires français (ISO/CEI 14651). • La recherche devra se faire par mots clés. Elle ne doit pas tenir compte de la casse ni de l'accentuation.

2.3.3. Adaptabilité culturelle et linguistique dans les applications**Contexte**

Les applications informatiques gouvernementales destinées à l'usage d'un citoyen (individu) doivent prendre en compte la langue utilisée par ce citoyen, soit le français ou l'anglais, en conformité avec la Charte de la langue française. C'est le cas notamment du Service québécois de changement d'adresse (SQCA).

Pour répondre à ce besoin, il importe de mettre en œuvre le concept d'adaptabilité culturelle et linguistique¹ lors du développement d'une application. Ce point est traité dans le standard SGQRI 50 – *Gestion des données dans une application à portée multilingue*.

¹ Voir la définition de *adaptabilité culturelle et linguistique* en annexe.

À titre d'exemples, voici les éléments relatifs au traitement de l'information qui varient selon les caractéristiques des langues et de leurs rapports entre elles : caractères nationaux et éléments connexes (comme les traits d'union, les tirets et les signes de ponctuation), systèmes d'écriture, transformation appropriée des caractères, dates, poids et mesures, règles de tri et de recherche, codage des entités nationales (comme les codes de pays et de devise), présentation des numéros de téléphone et disposition des claviers.

Orientation

Les M/O doivent se référer au standard SGQRI 50 – *Gestion des données dans une application à portée multilingue*.

No.	Composant	Norme ou standard	Orientation	Références	Commentaires
s125	Adaptabilité culturelle et linguistique dans ldes applications	SGQRI 50	Prendre en compte	SGQRI 50 Communiquer avec Yves Hudon, Yves.hudon@msg.gouv.qc.ca	À titre d'exemples, voici les éléments relatifs au traitement de l'information qui varient selon les caractéristiques des langues et de leurs rapports entre elles : caractères nationaux et éléments connexes (comme les traits d'union, les tirets et les signes de ponctuation), systèmes d'écriture, transformation appropriée des caractères, dates, poids et mesures, règles de tri et de recherche, codage des entités nationales (comme les codes de pays et de devise), présentation des numéros de téléphone et disposition des claviers.

3. Soutien de l'internationalisation pour les données vocales

3.1 Identification de la langue utilisée dans un document destiné à la restitution vocale

Contexte

La restitution vocale à partir d'un texte doit prendre en compte le contexte linguistique et au minimum la langue utilisée.

L'identification de la langue utilisée dans un document destiné à la restitution vocale doit se faire idéalement à l'aide du codet alpha-3 (au besoin, voir le standard SGQRI 46-4 – *Codification de l'information d'un domaine énumératif codifié*, partie 4 : « Langues »). Des contraintes actuelles sur Internet nécessitent toutefois que l'on utilise le codet alpha-2 (qui ne permet que 676 langues sur les 3000 langues vivantes) si la restitution vocale implique l'usage d'Internet.

Ces codets décrivant la langue comportent trois ou au minimum deux caractères (exemple : *fra* ou *fr* pour le français) et utilisent l'ensemble de caractères ASCII américain (*a-z*). Malgré les limitations de cet ensemble (toutes les langues ne sont pas, et de loin, représentées), le codet alpha-2 (à deux caractères) est utilisé par le W3C.

Orientations

Les M/O doivent retenir le codet alpha-3, comme l'indique le standard SGQRI 46-4, pour identifier la langue utilisée dans un texte destiné à la restitution vocale, sauf dans les cas où Internet est utilisé, pour lesquels le codet alpha-2 est pour l'instant imposé.

No.	Composant	Norme ou standard	Orientations	Références	Commentaires
s130	Identification de la langue utilisée dans un document destiné à la restitution vocale	Codet alpha-3 du SGQRI 46-4	Prendre en compte	Standard en cours d'approbation Communiquer avec Yves Hudon, Yves.hudon@msg.gouv.qc.ca	- Les codets décrivant la langue comportent trois caractères (exemple : <i>fra</i> pour le français) et utilisent l'ensemble de caractères ASCII américain (<i>a-z</i>). - Malgré les limitations de cet ensemble (toutes les langues ne sont pas, et de loin, représentées), le codet alpha-2 est utilisé par le W3C (qui en fait se réfère au RFC 1766).

3.2 Reconnaissance de la parole

Contexte

Actuellement, la reconnaissance de la parole est un domaine en friche dont les applications technologiques restent encore très limitées.

De nombreuses recherches visent à modéliser et à développer un système de transcription de la parole performant quels que soient le locuteur et la langue parlée. Les performances d'un tel système dépendent notamment de la qualité des modèles acoustiques et linguistiques utilisés.

Le CNRS/LIMSI (*Laboratoire d'informatique pour la mécanique et la science de l'ingénieur du Centre national de la recherche scientifique*) (<http://www.limsi.fr/RS98FF/CHM98FF/TLP98FF/ChapTlp/node13.html>) explique en substance que ces modèles doivent rendre compte des phénomènes propres à la parole spontanée (hésitations, respirations, reprises, syntaxe de l'oral, ...), de la très grande variabilité acoustique (microphones, bruits de fond, transmission téléphonique, présence de musique, ...) ainsi que du contexte linguistique.

L'adaptation du système de reconnaissance à de nouvelles langues nécessite pour chaque langue un corpus² de parole, un corpus de textes et un lexique phonétique. Il peut également être nécessaire d'adapter la structure des modèles pour prendre en compte les phénomènes spécifiques à la langue tant sur le plan phonologique que syntaxique.

Des systèmes de reconnaissance de la parole continue à multilocuteurs ont été développés pour le français, l'anglais américain, l'anglais britannique et l'allemand, avec des vocabulaires de l'ordre de 64 000 mots.

Des travaux sont menés au sein du W3C notamment sur l'extraction et la traduction d'un texte à partir d'un outil de reconnaissance vocale s'appuyant sur une grammaire ([Semantic Interpretation](#)).

Orientation

Si des besoins sont exprimés par les M/O, des orientations visant la normalisation pourront être établies.

No.	Composant	Norme ou standard	Orientation	Références	Commentaires
s140	Internationalisation de la reconnaissance vocale	Aucun actuellement	Éligible dans une version ultérieure	Aucune actuellement	Domaine en friche.

² Voir la définition de *corpus* en annexe.

3.3 Synthèse de la parole à partir de texte (systèmes *text-to-speech*)

Contexte

Différentes technologies de synthèse de la parole existent. En ce domaine comme ailleurs, l'interopérabilité doit être favorisée en visant la conformité avec des standards ou des normes.

Comme il est indiqué dans le volet 2, c'est le composant e130 qui effectue l'adaptation et la présentation à l'utilisateur d'un document vocal en fonction de son profil. La recommandation [VoiceXML 2](#) rédigée par le W3C (voir le composant e130 dans le volet 2) est à prendre en compte, et plus particulièrement la spécification [SSML 1.0](#) (*Speech Synthesis Markup Language*), qui synthétise la parole à partir d'un texte balisé respectant le format SSML.

L'internationalisation de ce composant est prise en compte dans la spécification du W3C elle-même et intervient dans les aspects suivants :

- le document texte est un document utilisant la syntaxe SSML (donc conforme à la syntaxe XML) ;
- le contexte linguistique est pris en compte à différents niveaux de granularité en utilisant l'attribut `xml:lang`, qui vient caractériser les éléments « *voice* », « *speak* », « *paragraph* » et « *sentence* » ;
- le document texte peut référencer un ou plusieurs documents de prononciation lexicale et recourir à la définition d'une prononciation ou faire les deux à la fois, par le moyen de l'élément « *phoneme* ». Cette définition peut utiliser par exemple l'alphabet phonétique international³ soutenu notamment par le jeu universel de caractères (norme internationale ISO/CEI 10646, ou standard Unicode) ;
- des nuances de prosodie⁴ et de style sont apportées au travers des attributs optionnels « *gender* » (pouvant prendre les valeurs « *male* », « *female* », « *neutral* »), « *age* », « *variant* » et « *name* ».

Actuellement, il existe divers produits commerciaux qui supportent plusieurs langages au travers de SSML, mais la qualité n'est pas encore au rendez-vous.

Orientation

Les M/O peuvent reproduire vocalement un texte dans plusieurs langues en exploitant le potentiel de VoiceXML 2, et en particulier le module de synthèse vocale et son langage de description SSML 1.0. Ce standard est très prometteur, mais demande une meilleure assise dans l'industrie.

³ Voir la définition en annexe.

⁴ Voir la définition de *prosodie* en annexe.

No.	Composant	Norme ou standard	Orientation	Références	Commentaires
s150	Internationalisation de la synthèse vocale	SSML 1.0	À prendre en compte	http://www.w3.org/TR/speech-synthesis/	- Le document texte à reproduire vocalement est un document utilisant la syntaxe SSML version 1.0 (donc conforme à la syntaxe XML). - Le contexte linguistique est pris en compte à différents niveaux de granularité en utilisant l'attribut <code>xml:lang</code>, qui vient caractériser les éléments « <i>voice</i> », « <i>speak</i> », « <i>paragraph</i> » et « <i>sentence</i> ». - Le document SSML peut référencer un ou plusieurs documents de prononciation lexicale et/ou peut recourir à la définition d'une prononciation par le moyen de l'élément « <i>phoneme</i> ». Cette définition peut utiliser, par exemple, l'alphabet phonétique international. - Des nuances de prosodie et de style sont apportées au travers des paramètres optionnels « <i>gender</i> » (pouvant prendre les valeurs « <i>male</i> », « <i>female</i> », « <i>neutral</i> »), « <i>age</i> », « <i>variant</i> » et « <i>name</i> ».

Annexe : définitions

Adaptabilité culturelle et linguistique	Selon le comité mixte sur les technologies de l'information de l'Organisation internationale de normalisation (ISO) et de la Commission électrotechnique internationale (ISO/CEI JTC 1), l'adaptabilité culturelle et linguistique représente la capacité d'un produit à satisfaire aux exigences suivantes tout en maintenant sa portabilité et son interopérabilité : • pouvoir être internationalisé, c'est-à-dire pouvoir être adapté tant aux caractéristiques spéciales des langues et à leurs règles générales d'emploi qu'aux cultures d'une région géographique donnée ; • pouvoir tenir entièrement compte des besoins d'une catégorie d'utilisateurs.
Alphabet phonétique international	International Phonetic Association. <i>International Phonetic Alphabet</i>, Victoria (British Columbia), Department of Linguistics, University of Victoria, 1996. Voir <http://www2.arts.gla.ac.uk/IPA/fullchart.html>. Voir aussi la référence <http://sivanataraja.free.fr/api/>.
Corpus	Ensemble des sources orales et écrites relatives au domaine étudié et qui sont utilisées dans un travail terminologique (source : Office québécois de la langue française, <i>Le grand dictionnaire terminologique</i>).
Prosodie	Ensemble des éléments phoniques (intonation affective, particularismes régionaux, accent tonique, montée mélodique, etc.) qui caractérisent le langage parlé (source : Office québécois de la langue française, <i>Le grand dictionnaire terminologique</i>).

*Secrétariat
du Conseil du trésor*

Québec

