Network Working Group
Request for Comments: 2070
Category: Standards Track

F. Yergeau
Alis Technologies
G. Nicol
Electronic Book Technologies
G. Adams
Spyglass
M. Dürst
Université de Zurich

Janvier 1997

Internationalisation du langage de balisage hypertexte (HTML)

Position de ce document

Ce document fournit à la communauté Internet un protocole Internet en voie de normalisation, et appelle à la discussion et aux commentaires en vue de son amélioration. Veuillez consulter l'édition courante de Internet Official Protocol Standards (STD 1) pour le niveau de normalisation et la position de ce protocol. La distribution de ce document n'est pas restreinte.

Résumé

Le langage de balisage hypertexte (HTML) est un langage de balisage utilisé pour créer des documents hypertexte indépendant de tout système. À l'origine, l'application d'HTML sur le World Wide Web était sérieusement limitée par son utilisation du jeu de caractères codés ISO-8859-1, utile seulement pour les langues d'Europe occidentale. Malgré cette restriction, HTML a été utilisé en d'autres langues, avec d'autres jeux ou codages de caractères, aux dépens de l'interopérabilité.

Ce document porte sur l'internationalisation (i18n, i suivi de 18 lettres suivies de n) d'HTML ; il élargit la spécification d'HTML et donne des conseils additionnels pour une gestion correcte de l'internationalisation. Une considération de premier plan est de s'assurer qu'HTML demeure une application valide de SGML, tout en permettant son utilisation avec toutes les langues du monde.

Table des matières

---

1. Introduction

Le langage de balisage hypertexte (HTML) est un langage de balisage utilisé pour créer des documents hypertexte indépendant de tout système. À l'origine, l'application d'HTML sur le World Wide Web était sérieusement limitée par son utilisation du jeu de caractères codés ISO-8859-1, utile seulement pour les langues d'Europe occidentale. Malgré cette restriction, HTML a été utilisé en d'autres langues, avec d'autres jeux ou codages de caractères, par le biais de diverses extensions ad hoc au langage [TAKADA].

Ce document porte sur l'internationalisation d'HTML ; il élargit la spécification d'HTML et donne des conseils additionnels pour une gestion correcte de l'internationalisation. Il est en bonne part basé sur un article d'un des auteurs sur le multilinguisme sur le WWW [NICOL]. Une considération de premier plan est de s'assurer qu'HTML demeure une application valide de SGML, tout en permettant son utilisation avec toutes les langues du monde.

Les principaux sujets traités sont le jeu de caractères de document à utiliser avec HTML, le traitement correct du paramètre charset associé au type de contenu text/html et la spécification de quelques éléments et entités supplémentaires.

1.1 Domaine d'application

HTML est utilisé sur le système mondial d'information World Wide Web (WWW) depuis 1990. Ce document étend les capacités d'HTML 2.0 (RFC 1866), principalement en enlevant la restriction au jeu de caractères codés ISO-8859-1 [ISO-8859].

HTML est une application de la norme ISO 8879:1986, Traitement de l'information — Systèmes bureautiques — Langage normalisé de balisage généralisé (SGML) [ISO-8879]. La Définition de Type de Document (DTD) d'HTML est une définition formelle de la syntaxe HTML en termes SGML. Ce document modifie la DTD d'HTML 2.0 de façon à la rendre applicable à des documents comprenant un répertoire de caractères beaucoup plus grand que celui de l'ISO-8859-1, tout en conservant la conformité avec SGML.

Le développement d'HTML avance très vite, autant formellement que pratiquement. Ce document est écrit de manière à ce que les changements préconisés à HTML puissent (et devraient) s'appliquer à d'autres formes d'HTML que celle décrite dans le RFC 1866. Lorsque indiqué, les nouveaux attributs devraient s'appliquer au éléments appropriés.

1.2 Conformité

Cette spécification change légèrement les exigences de conformité pour les documents et agents-usager HTML.

1.2.1 Documents

Tous les documents conformes à HTML 2.0 demeurent conformes. Toutefois, les extensions introduites ici rendent valides certains documents qui ne seraient pas conformes à HTML 2.0, en particulier ceux contenant des caractères ou des références de caractères hors du répertoire de l'ISO 8859-1, et ceux contenant du balisage nouveau.

1.2.2. Agents-usager

En sus des exigences du RFC 1866, les exigences suivantes s'appliquent aux agents-usager HTML.

- Pour assurer l'interoperabilité et une gestion correcte d'au moins l'ISO-8859-1 dans un environnement où des encodages de caractères autres que l'ISO-8859-1 sont présents, les agents-usager DOIVENT interpréter correctement le paramètre charset qui accompagne un document reçu du réseau.
- De plus, les agents-usager conformes DOIVENT au moins analyser correctement toute référence numérique de catactères dans le domaine de l'ISO 10646-1 [ISO-10646].
- Les agents-usager conformes doivent appliquer l'algorithme de présentation BIDI s'il affichent des caractères de droite-à-gauche. Il n'y a pas lieu d'appliquer de traitement BIDI si un document ne contient aucun caractère droite-à-gauche affichable.

---

2. Le jeu de caractères de document

2.1. Modèle de traitement de référence

Cet aperçu explique un modèle de traitement de référence pour HTML, et en particulier le concept SGML de jeu de caractère de document. Une quelconque mise en oeuvre peut être très différente du modèle, mais devrait se comporter de la même façon pour un observateur externe.

Étant donné l'existence d'une grande variété de codages de texte, SGML ne s'intéresse pas directement à la façon dont les séquences de caractères constituant un document SGML au sens abstrait sont codés en séquences d'octets (ou parfois de groupes de bits de longueur différente de 8), lors d'une réalisation concrète du document comme un fichier. Ce codage est appelé le codage externe de caractères du document SGML concret, et doit être soigneusement distingué du du jeu de caractères de document du document SGML abstrait. SGML considère un unique ensemble de caractères (appelé un répertoire de caractères), et un jeu de codes qui associe à chaque caractère du répertoire un nombre entier (appelé le numéro du caractère). La déclaration du jeu de caractères de document définit ce que chaque numéro de caractères représente [GOLD90, p. 451]. Dans la plupart des cas, une DTD SGML et tous les documents qui s'y réfèrent ont un seul jeu de caractères de document, et tout le balisage et les caractères de texte en font partie.

HTML, en tant qu'application SGML, ne s'inquiète pas du codage externe de caractères. Ce problème est délégué à des mécanismes externes, tels que MIME dans le cas du protocole HTTP ou du courrier électronique.

Le protocole HTTP [RFC2068] utilise le paramètre charset du champ Content-Type de l'en-tête de la réponse pour indiquer le codage externe de caractères. Par exemple, pour indiquer qu'un document japonais est codé en JUNET [RFC1468], l'en-tête contiendra la ligne suivante :

```
Content-Type: text/html; charset=ISO-2022-JP
```

L'expression MIME charset est utilisée pour désigner un codage de caractères, et non pas seulement un jeu de caractères codés tel que l'expression semble l'indiquer. Un codage de caractères est une relation (possiblement de plusieurs à un) entre une suite d'octets et une suite de caractères tirés d'un ou de plusieurs répertoires.

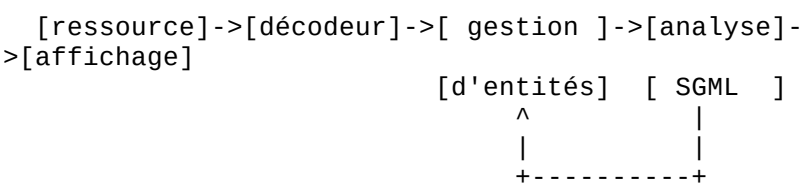
Le protocole HTTP définit aussi un mécanisme pour que le client puisse indiquer les codages qu'il accepte. Les serveurs et clients sont fortement encouragés à utiliser ces mécanismes pour assurer la transmission et l'interprétation correctes de tout document. Des mesures compensatoires, à utiliser quand les mécanismes corrects sont ignorés, sont décrits à la section 6.

De même, lorsqu'un document HTML est transmis par courrier électronique, le codage externe de caractères est précisé par le paramètre charset de la ligne d'en-tête MIME Content-Type [RFC2045], avec US-ASCII comme valeur implicite.

Il n'existe présentement aucun mécanisme normalisé pour indiquer le codage externe de caractères de documents HTML transmis par FTP ou obtenus de systèmes de fichiers distribués.

Au cas où tout autre moyen de transmettre ou de stocker des documents HTML soit défini ou devienne populaire, il est conseillé d'avoir de semblables moyens d'identification du codage, ou encore d'utiliser ou d'avoir par défaut un codage capable de représenter un grand nombre de caractères en contexte international.

Quelque soit le codage de caractères externe, le modèle de traitement de référence le traduit au jeu de caractères de document spécifié à la section 2.2 avant tout traitement propre à SGML/HTML. Le modèle de traitement de référence peut être illustré comme suit :



Le décodeur est responsable du décodage de la représentation externe vers le jeu de caractères de document. Le gestionnaire d'entités, l'analyseur et l'application ne voient que des caractères du jeu de caractères de document. Le mécanisme d'affichage, ou encore la partie affichage de l'application, peut encore recoder les caractères en une représentation plus appropriée à sa fonction. Dans tous les cas, et en tout ce qui concerne la sémantique des caractères, le

gestionnaire d'entités, l'analyseur et l'application utilisent exclusivement le jeu de caractères de document.

Une quelconque mise en oeuvre peut choisir, ou non, de traduire le document en une représentation du jeu de caractères de document tel que décrit ci-dessus ; le comportement de ce modèle de traitement de référence peut être obtenu autrement. Ce sujet est toutefois hors du domaine de cette spécification, et le lecteur est invité à consulter la norme SGML [ISO-8879] ou un manuel SGML [BRYAN88] [GOLD90] [VANH90] [SQ91] pour plus ample informé.

La conséquence la plus importante de ce modèle de référence est que les références numériques de caractères sont toujours relatives au jeu de caractères de document, sans égard au codage externe du document. Voir la section 2.2 pour un exemple.

2.2. Le jeu de caractères de document

Le jeu de caractères de document, au sens SGML, est le jeu universel de caractères (JUC) de l'ISO 10646:1993 [ISO-10646], tel qu'amendé. À l'heure actuelle, ce jeu est identique à celui du standard Unicode, version 1.1 [UNICODE].

> NOTE — les implémenteurs devraient savoir que l'ISO 10646 est amendée de temps en temps ; 4 amendements ont été adoptés depuis la parution initiale de 1993, sans que cette spécification n'en soit affectée significativement. Un cinquième amendement, présentement à l'étude, amènera des changements incompatible à la norme : 6556 syllabes coréennes Hangul ayant des valeurs de code entre 3400 et 4DFF (hexadécimal) seront déplacées à de nouvelles valeurs (avec ajout de 4516 nouvelles syllabes), rendant ainsi les références aux anciennes valeurs invalides. Puisque le consortium Unicode a déjà adopté un amendement à cet effet pour Unicode 2.0 (à paraître bientôt), l'adoption du DAM 5 est probable et les implémenteurs devraient probablement déjà considérer les anciennes valeurs comme obsolètes. Malgré ce changement unique, les organismes de normalisation pertinents semblent demeurer opposés à tout changement de valeurs de code dans l'avenir. Pour coder le coréen sans égard à ce changement, on peut utiliser les Jamo Hangul combinatoires situés de 1110 à 11F9.

L'adoption de ce jeu de caractères de document impose un changement à la déclaration SGML de HTML 2.0 (section 9.5 de [RFC1866]). Il s'agit de remplacer la première déclaration BASESET et le DESCSET correspondant par ce qui suit :

```
BASESET "ISO Registration Number 177//CHARSET
         ISO/IEC 10646-1:1993 UCS-4 with implementation level 3
         //ESC 2/5 2/15 4/6"
DESCSET  0   9     UNUSED
         9   2     9
         11  2     UNUSED
         13  1     13
         14  18    UNUSED
         32  95    32
         127 1     UNUSED
         128 32    UNUSED
         160 2147483486 160
```

L'adoption du JUC comme jeu de caractère de document conserve la conformité de toute expression, toute construction ou tout document conforme à HTML 2.0. Elle rend toutefois

conforme certaines contructions non-conforme à HTML 2.0. En particulier, les caractères hors du répertoire de l'ISO 8859-1 mais dans le répertoire de l'UCS-4 deviennent valides, et la borne supérieure des références numériques de caractères est portée de 255 à 2147483645 ; ainsi, И est une référence valide à la LETTRE MAJUSCULE CYRILLIQUE I (?). [ERCS] est une bonne source de renseignements sur Unicode et SGML, bien que son domaine et son contenu technique soient très différents de ceux de ce document.

> NOTE — la déclaration SGML ci-dessus, comme celle d'HTML 2.0, donne les caractères de 128 à 159 (80 to 9F hex) comme inutilisés (UNUSED). Ceci signifie que les références numériques de caractères correspondantes (par ex. ’) sont illégales en HTML. Ni ISO 8859-1 ni ISO 10646 ne contiennent de caractères dans ce domaine, qui est réservé à des caractères de contrôle.

Une autre différence avec la déclaration SGML de HTML 2.0 est due à l'opinion que cette dernière n'exprime pas vraiment l'intention de ses auteurs. La déclaration du jeu de caractères de syntaxe passe de ISO 646.IRV:1983 au plus récent ISO 646.IRV:1991, ce dernier étant identique à US-ASCII, contrairement au premier. En principe, ceci introduit une incompatibilité avec HTML 2.0, mais en pratique l'interopérabilité devrait être améliorée par i) une déclaration SGML conforme à ce que tout le monde pense, et ii) un jeu de caractères de syntaxe sous-ensemble propre du jeu de caractères de document. Les caractères qui diffèrent entre les deux versions de l'ISO 646.IRV ne sont pas utilisés pour exprimer la syntaxe d'HTML.

L'ISO 10646-1:1993 est le plus grand jeu de caractères qui soit, et aucun autre jeu de caractères ne pourrait le remplacer comme jeu de caractères de document pour HTML. Toutefois, si pour une certaine application il s'avérait nécessaire d'utiliser d'autres caractères, ce devrait être fait de manière compatible avec les versions présente et futures d'ISO 10646, i.e. en plaçant ces caractères en zone privée de l'espace de codage UCS-4 [ISO-10646, section 11]. On doit aussi se rendre compte qu'un tel usage serait très peu portable ; il vaudra généralement mieux utiliser de petites images en ligne.

2.3. Caractères non-affichables

Avec l'ISO 10646 au complet comme jeu de caractères de document, on ne peut éviter le problème de caractères qui ne peuvent être affichés faute de ressources appropriées (fontes). Comme il y a nombre de solutions possibles, y compris de laisser le mécanisme d'affichage sous-jacent se débrouiller, ce document ne prescrit pas de comportement particulier. Les conseils suivants pourront toutefois être utiles :

- On préfèrera un comportement bien visible mais non nuisible. Comme certains documents pourraient contenir un grand nombre de caractères non-affichables, une alerte pour chacun n'est pas appropriée.
- Si une représentation numérique du caractère est montrée, la forme hexadécimale devrait être affichée de préférence à la forme décimale, pour une meilleure correspondance avec les normes de jeux de caractères [ERCS].

---

3. L'attribut LANG

Les étiquettes de langue peuvent être utiles pour différents aspects du rendu de documents balisés : choix de glyphe, dans les cas d'ambiguïté non résolues par le codage de caractères ; césure ; ligatures ; espacement ; synthèse vocale ; etc. Indépendamment des questions de rendu, l'étiquetage de langue est utile comme balisage de contenu pour des fins comme la classification et la recherche.

Puisque qu'à tout texte peut logiquement être associé une langue, presque tous les éléments HTML admettent l'attribut LANG, ainsi qu'en témoigne la DTD; les seuls éléments de cette version d'HTML à ne pas admettre l'attribut LANG sont BR, HR, BASE, NEXTID et META. L'intention est aussi que tout nouvel élément d'une version ultérieure d'HTML admette l'attribut LANG, sauf contre-indication.

L'attribut de langue, LANG, a comme valeur une étiquette qui identifie une langue naturelle parlée, écrite ou autrement exprimée par des personnes en vue de communiquer avec d'autres personnes. Les langages de programmation sont explicitement exclus.

La syntaxe et le registre des étiquettes de langue HTML sont les mêmes que ceux définis par le RFC 1766 [RFC1766]. En résumé, une étiquette de langue se compose d'une ou de pluieurs parties : une étiquette primaire et une série, pouvant être vide, de sous-étiquettes :

```
étiquette-de-langue = étiquette-primaire *( "-" sous-étiquette )
étiquette-primaire  = 1*8ALPHA
sous-étiquette      = 1*8ALPHA
```

Il n'y a pas de blanc dans une étiquette, et elle sont insensibles à la casse. L'espace des noms des étiquettes de langue est administré par l'IANA. Exemples :

```
fr, fr-FR, en-cockney, i-cherokee, x-latin-de-cuisine
```

En HTML, une étiquette de langue ne doit pas être interprétée comme indivisible, selon le RFC 1766, mais comme une hiérarchie. Par exemple, un agent-usager qui règle le rendu selon la langue devrait conclure qu'il y a correspondance lorsque qu'une étiquette de langue dans une feuille de style correspond à la portion initiale de l'étiquette de langue d'un élément, à défaut d'une correspondance exacte. Cette interprétation permet à un élément balisé comme, par exemple, fr-FR d'utiliser des styles prévus pour le français de France (fr-FR) ou à défaut le français générique ou international (fr).

> NOTE — L'interprétation en hiérarchie des étiquettes de langue ne signifie pas que toutes les langues ayant un préfixe commun sont comprises par les locuteurs de l'une ou plusieurs de ces langues ; on permet simplement aux utilisateurs d'en profiter lorsque c'est le cas.

Le rendu des éléments peut être affecté par l'attribut LANG. Pour tout élément, la valeur de son attribut LANG l'emporte sur la valeur de l'attribut LANG d'un élément englobant et sur la valeur (s'il y a lieu) de l'en-tête HTTP Content-Language. À défaut, une valeur implicite, possiblement soumise à réglage par l'utilisateur, devrait contrôler le rendu.

---

4. Entités, attributs et éléments supplémentaires

4.1. Jeu d'entités Latin-1 complet

D'après la suggestion de la section 14 du [RFC1866], le jeu d'entités Latin-1 est augmenté pour comprendre toute la partie droite de l'ISO 8859 (toutes les positions dont le bit de poids fort est 1), y compris , © (©) et ® (®) déjà fortement utilisés. Les noms des entités proviennent des appendices de SGML [ISO-8879]. On en trouvera la liste à la section 7.3 de ce document.

4.2. Balisage pour présentation dépendant de la langue

4.2.1. Survol

Pour le rendu correct de texte en certaines langues (sans égard au formattage), certaines entités et certains éléments additionnels sont nécessaire.

En particulier, on s'intéressera au cas suivants :

- Balisage de texte bidirectionnel, c'est à dire de texte où des écritures de droite à gauche et de gauche à droite sont présentes ;
- Contrôle du comportement de ligature cursive, dans les cas où le comportement implicite n'est pas adéquat ;
- Rendu dépendant de la langue de courtes citations (en ligne) ;
- Meilleure contrôle sur la composition des paragraphes pour les langues qui l'exigent ;
- Indices et exposants pour les langues où ils apparaissent au sein de texte ordinaire.

Certains de ces cas n'ont besoin que de très peu de soin ; d'autres sont plus exigeants. Les caractéristiques supplémentaires ne sont introduites ici qu'avec de brefs commentaires. Des explications sur les textes cursifs et/ou bidirectionnel suivent. Dans ces deux derniers cas, ce document suit [UNICODE] en ce que : i) la sémantique des caractères, lorsqu'applicable, est identique à celle d'[UNICODE], et ii) quand une fonction est transposée vers HTML en tant que protocole de niveau supérieur, la transposition est faite de manière à permettre une conversion immédiate aux mécanismes de bas niveau définis dans [UNICODE].

4.2.2. Liste des entités, éléments et attributs

D'abord, on a besoin d'un conteneur générique pour porter les attributs LANG et DIR (voir ci-bas) dans les cas ou aucun autre élément n'est approprié l'élément SPAN est introduit à cet effet.

Un jeu d'entités de caractères nommées est ajouté pour le rendu bidirectionnel et le contrôle des ligatures cursives :

```
<!ENTITY zwnj CDATA "&#8204;" -- =anti-liant sans chasse -->
<!ENTITY zwj  CDATA "&#8205;" -- =liant sans chasse -->
<!ENTITY lrm  CDATA "&#8206;" -- =marque gauche-à-droite -->
<!ENTITY rlm  CDATA "&#8207;" -- =marque droite-à-gauche -->
```

Ces entités peuvent être utilisées en lieu et place des caractères de formattage correspondant lorsque pratique, par exemple pour faciliter la saisie ou lorsqu'un caractère de formattage n'est pas disponible dans le codage de caractères du document.

Ensuite, un attribut appelé DIR est introduit, restreint aux valeurs LTR (gauche à droite) et RTL (droite à gauche) et admis par la plupart des éléments, servant à indiquer la directionnalité en contexte bidirectionnel (voir 4.2.4 ci-bas pour détails). Puisque la

directionnalité s'applique logiquement à tout texte et à plusieurs autres éléments (par ex. les tableaux), tous les éléments sauf BR, HR, BASE, NEXTID et META admettent cet attribut, tel que l'indique la DTD. À moins d'une bonne raison, tout nouvel élément introduit dans une version ultérieure d'HTML devrait aussi admettre cet attribut.

Un nouvel élément en-ligne appelé BDO (surcharge Bidi) est introduit, exigeant l'attribut DIR pour indiquer si la surcharge est de gauche-à-droite ou de droite-à-gauche. Cet élément est nécessaire pour le contrôle de texte bidirectionnel, voir les détails à la section 4.2.4.

L'élément en-ligne Q est introduit pour permettre le rendu dépendant de la langue de citations courtes selon la langue et les capacités du système. Comme le montrent les exemples suivants, les guillemets entourant la citation sont particulièrement affectés : "une citation en anglais", `une autre, un peu mieux', ,,une citation en allemand", « une citation en français ». Le contenu de l'élément Q n'inclut pas les guillemets, qui doivent être ajoutés au moment du rendu.

> NOTE — Les éléments Q peuvent s'imbriquer. En plusieurs langues, le style des citations varie selon l'imbrication, ce que les agents-usager devrait prendre en compte.

> NOTE — une gestion minimale de l'élément Q consiste à entourer le contenu de quelconques guillemets, même les simples guillemets doubles ASCII. Étant donné la facilité de réalisation, et étant donné la possible perte de sens en l'absence de toute marque visible, on encourage fortement les programmeurs d'agents-usager à inclure au moins ce minimum.

Plusieurs langues exigent des exposants pour un rendu correct : par exemple, en français le lle de Mlle Dupont devrait être en exposant. L'élément SUP, et son cousin SUB pour les indices, sont introduits pour permettre le balisage de tel texte. Le contenu de SUP et SUB est restreint à PCDATA pour éviter des problèmes d'imbrication.

Finalement, en plusieurs langues la justification a beaucoup plus d'importance qu'en langues occidentales, et justifie du balisage. L'attribut ALIGN, admettant les valeurs LEFT, RIGHT, CENTER et JUSTIFY, est ajouté à quelques éléments pour lesquels la justification a du sens (les éléments de type bloc P, HR, H1 to H6, OL, UL, DIR, MENU, LI, BLOCKQUOTE et ADDRESS). Un agent-usager qui choisit LEFT comme valeur implicite pour des blocs de directionnalité gauche-à-droite devrait utiliser RIGHT pour des blocs droite-à-gauche.

> NOTE — la section 4.2.2 du RFC 1866 indique qu'un agent-usager HTML devrait traiter une fin de ligne comme une espace de mot, sauf au sein de texte préformaté. Ceci doit être interprété dans le contexte de l'écriture traitée, puisque différentes écritures ont différentes manières de séparer les mots : une simple espace dans certaines écritures (par ex. latine), un séparateur de mot sans chasse dans d'autres (par ex. thaï), et rien du tout (totalement ignorée) dans certaines autres (par ex. japonaise).

> NOTE — les implémenteurs d'agents-usager devraient porter une attention spéciale au TRAIT D'UNION VIRTUEL (U+00AD), un caractère qui se retrouve dans de nombreux jeux de caractères (dont toute la série ISO 8859 et, bien sûr, l'ISO 10646), qui peut toujours être saisi comme une référence , et qui est différent du simple TRAIT D'UNION : il indique où un bris de ligne

est permis. Si la ligne est effectivement brisée, un trait d'union doit être affiché à la fin de la première ligne. Sinon, le caractère n'est aucunement affiché. Dans des opération comme le tri et la recherche, il devrait être complètement ignoré.

Dans la DTD, les attributs LANG et DIR sont regroupés dans une entité de paramètre appelée *attrs*. À l'instar du RFC 1942 [RFC1942], les attributs ID et CLASS sont aussi groupés dans *attrs*. Les attributs ID et CLASS sont exigés par les feuilles de style, et le RFC 1942 les décrit comme suit :

ID
: Utilisé pour définir un identifieur au niveau du document. Peut être utilisé pour nommer une position dans un document comme cible d'un hyperlien. Peut aussi être utilisé par des feuilles de style pour rendre un élément dans un style particulier. Un attribut ID est un NOM SGML un tel NOM est formé d'au moins une lettre suivie de lettres, de chiffres, de « - » ou de « . ». Les lettres sont limitées à A-Z et a-z.

CLASS
: Une liste de NOMs SGML séparés par des espaces. Les noms de classes indiquent que l'élément appartient aux classes correspondantes, ce qui permet aux auteurs de distinguer différents rôles pour le même élément. Les classes peuvent être utilisées par des feuilles de style pour offrir un rendu différent selon les rôles.

4.2.3. Ligature cursive

On a parfois besoin de balisage pour imposer une ligature cursive dans des cas où elle ne se produirait normalement pas, ou pour l'empêcher quand elle se produrait.

Les liant et anti-liant sans chasse (‍ et ‌) sont utilisés pour contrôler les ligatures cursives. Par exemple, la LETTRE ARABE HA' est utilisée seule comme abbréviation de Hijri (le calendrier islamique) ; toutefois, on veut la forme initiale de la lettre (ﻫـ), la forme isolée de HA' (ﻩ) ressemblant trop au chiffre cinq de l'écriture arabe. On obtient cet effet en mettant un liant sans chasse à la suite de HA', dont le seul effet est de procurer du contexte. Dans des textes persans, il arrive qu'une lettre ne se joigne pas à la suivante alors qu'elle le ferait normalement. On utilise ici un anti-liant sans chasse.

4.2.4. Texte bidirectionnel

De nombreuses langues s'écrivent en lignes horizontales de gauche à droite, alors que d'autres s'écrivent de droite à gauche. Lorsque les deux directions d'écriture sont présentes, on parle de texte bidirectionnel (BIDI en bref). Le texte BIDI exige du balisage lorsque des ambiguïtés quant à la directionnalité de caractères doivent être levées. Ce balisage permet de rendre du texte BIDI de façon sémantiquement lisible ; sans ce balisage BIDI spécial, certains cas ne peuvent être rendu **d'aucune** manière qui reflète le sens du texte. Le texte simple (non-enrichi) peut contenir du balisage BIDI sous la forme de caractères de formatage spéciaux.

Ceci est aussi possible en HTML, qui admet les cinq caractères de formatage de l'ISO 10646 reliés au BIDI (202A - 202E). Comme alternative, HTML offre du balisage SGML équivalent.

Le BIDI est un sujet complexe, et la conversion de séquences de texte en ordre logique en séquences d'affichage doit être faite selon l'algorithme et les propriétés de caractères précisés par [UNICODE]. Nous ne fournissons ici que juste assez d'explications pour faire comprendre la nécessité du balisage et pour en définir le sens exact.

L'algorithme BIDI Unicode est basé sur l'enregistrement des caractères du texte en ordre logique, qui est l'ordre normal de saisie et aussi l'ordre normal d'élocution. Pour en rendre le rendu possible, l'algorithme attribue à chaque caractères une directionnalité, par ex. gauche-à-droite pour les lettres latines et droite-à-gauche pour les caractères arabes et hébreux.

Les marques gauche-à-droite et droite-à-gauche (‎ et ‏) sont utilisées pour lever l'ambiguïté sur la directionnalité des caractères neutres. Par exemple, lorsqu'un guillemet anglais double (") est coincé entre une lettre latine et une lettre arabe, sa direction est ambiguë ; si une marque directionnelle est ajoutée de part ou d'autre, de façon à ce que le guillemet soit entouré de caractères d'une seule direction, l'ambiguïté est levée. Ces caractères sont comme des espaces sans chasse avec propriété directionnelle (mais sans propriété de brisure de mot ou de ligne).

Les enchâssures de texte contra-directionnelles imbriquées, due à des citations imbriquées ou à du collage de texte d'un contexte BIDI à un autre, sont un autre cas exigeant du balisage. De plus, il est fréquemment souhaitable d'indiquer la directionnalité de base d'un bloc de texte. Pour ces fins, l'attribut DIR est utilisé.

Sur les éléments de type bloc, l'attribut DIR indique la directionnalité de base du texte du bloc ; lorsqu'omis, la directionnalité est hérité du parent. La directionnalité implicite du document HTML complet est de gauche à droite.

Sur les éléments en ligne, DIR marque le début d'un nouveau niveau d'enchâssement (expliqué ci-bas) ; lorsqu'omis, l'élément en ligne ne démarre pas un nouvel enchâssement BIDI.

> NOTE — les éléments PRE, XMP and LISTING admettent l'attribut DIR, ce qui indique que le contenu ne doit pas être considéré comme pré-formatté quant à la disposition bidirectionnelle. L'algorithme BIDI doit encore être appliqué à chaque ligne.

Voici un exemple d'un cas où le balisage des enchâssure est requis, en montrant l'effet :

> Étant donné les lettres latines (en capitales) et arabes (en minuscule) suivantes en mémoire avec les enchâssures indiquées :
>
> <SPAN DIR=LTR> AB <SPAN DIR=RTL> xy <SPAN DIR=LTR> CD </SPAN> zw </SPAN> EF </SPAN>
>
> On obtient le rendu suivant, [] montrant les transitions directionnelles :
>
> [AB [wz [CD] yx] EF]
>
> Sans balisage et en contexte de gauche à droite, on aurait obtenu :
>
> [AB [yx] CD [wz] EF]
>
> Notez que le **yx** est à gauche et le **wz** à droite, à l'encontre du cas avec balisage explicite des enchâssures. Sans balisage, on n'a qu'au plus deux niveaux : un niveau directionnel de base et un seul niveau à contre-courant.

L'attribut DIR sur les éléments en ligne est équivalent aux caractères de formattage ENCHÂSSEMENT GAUCHE-À-DROITE (202A) et ENCHÂSSEMENT DROITE-À-GAUCHE (202B) de l'ISO 10646. La balise de fin est équivalente au caractère DÉPILEMENT DE FORMATAGE DIRECTIONNEL (202C).

La surchage directionnelle, fournie par l'élément BDO, est nécessaire pour de courts morceaux de texte dans lesquels la directionnalité ne peut être déterminée de façon non-ambiguë du seul contexte. Par exemple, on l'utilisera pour forcer l'affichage de gauche à droite (ou de droite à gauche) de numéros de pièces formés de lettres latines, de chiffres et de lettres hébraïques.

L'effet de BDO est d'imposer la directionnalité de tous les caractères à la valeur de DIR, sans égard à leurs propriétés directionnelles implicites. Il est équivalent à l'utilisation des caractères FORÇAGE GAUCHE-À-DROITE (202D) ou FORÇAGE DROITE-À-GAUCHE (202E) de l'ISO 10646, la balise de fin étant encore ici équivalente au caractère DÉPILEMENT DE FORMATAGE DIRECTIONNEL (202C).

> NOTE — les rédacteurs et auteurs de logiciels de rédaction noteront qu'il peut y avoir conflit si l'attribut DIR est utilisé sur des éléments en ligne (y compris BDO) en même temps que les caractères de formattage correspondants de l'ISO 10646.
>
> Il est préférable d'utiliser un ou l'autre exclusivement ; le balisage est plus à même de garantir l'intégrité structurelle du document, et pallie certains problèmes d'édition de texte bidirectionnel à l'aide d'un éditeur de texte simple, mais certains logiciels seront peut-être plus à l'aise avec les caractères de formattage. Si les deux méthodes sont utilisées, on doit prendre grand soin d'assurer l'imbrication correcte du balisage et des enchâssures ou surchages directionnelles, sans quoi les résultats au rendu sont indéfinis.

---

5. Formulaires

5.1. Ajouts à la DTD

Les formulaires étant un des seuls moyens d'obtenir des données de l'utilisateur, il est naturel de s'attendre à y trouver du texte en n'importe quelle langue. Bien que ce soit principalement un sujet d'interface-utilisateur, quelques considérations au niveau HTML sont utiles pour conseiller et promouvoir l'interopérabilité.

Pour assurer la pleine interopérabilité, il est nécessaire que l'agent-usager (et l'utilisateur) ait une indication du (ou des) codage de caractères que le serveur fournissant un formulaire saura accepter à la réception du formulaire rempli. Cette indication est fournie par le nouvel attribut ACCEPT-CHARSET des éléments INPUT et TEXTAREA, calqué sur l'en-tête HTTP Accept-Charset (voir [HTTP-1.1]), qui contient une liste de jeux de caractères, séparés par des espaces et/ou des virgules, acceptables par le serveur. Un agent-usager pourrait informer l'utilisateur du contenu de cet attribut, ou l'empêcher de saisir des caractères hors des répertoires des dits jeux de caractères.

NOTE — la liste de jeux de caractères doit être interprétée comme une liste OU-EXCLUSIF ; le serveur annonce que, pour chaque partie d'une entité multipartite, il peut accepter UN SEUL codage de caractères parmi la liste. Le client peut transcoder au besoin pour satisfaire le serveur.

NOTE — la valeur implicite de l'attribut ACCEPT-CHARSET d'un élément INPUT ou TEXTAREA est la valeur réservée UNKNOWN. On pourra interpréter cette valeur comme celle du codage de caractères utilisé par le document contenant cet élément.

5.2. Soumission de formulaires

Le mécanisme de soumission de formulaires d'HTML 2.0, fondé sur le type de contenu application/x-www-form-urlencoded, est pauvre en ce qui concerne l'internationalisation. En fait, les URL étant limités à l'ASCII, le mécanisme est inadéquat même pour du texte ISO-8859-1. La section 2.2 du [RFC1738] précise que des octets peuvent être codés selon la notation %HH, mais le texte d'un formulaire est composé de caractères, et non pas d'octets. Sans définition du codage de caractères, la notation %HH n'a pas de sens.

La meilleure solution est d'utiliser le type de contenu multipart/form-data, décrit dans le [RFC1867], avec la méthode POST de soumission de formulaires. Ce mécanisme emballe chaque paire nom-valeur dans une partie d'un corps multipartite MIME qui est transmis comme entité HTTP ; chaque partie peut être étiquetée avec un Content-Type approprié, y compris si nécessaire un paramètre charset qui spécifie le codage de caractères. Les changements à la DTD nécessaires pour cette méthode de soumission de formulaires ont été incorporés dans la DTD de ce document.

Une solution moins satisfaisante est d'ajouter un paramètre charset à l'étiquette de type de contenu application/x-www-form-urlencoded transmis avec une soumission de formulaire de type POST ; il est alors entendu que le codage URL du [RFC1738] est appliqué en sus du codage de caractères, en tant que Content-Transfer-Encoding implicite.

Il est regrettable que les fureteurs actuels ne permettent aux signets de préciser la méthode POST, ce qui devrait être corrigé. La méthode GET pourrait aussi être utilisée avec les données du formulaire comme entité plutôt que codé dans l'URL. Rien dans le protocole ne semble l'interdire, mais aucune réalisation ne semble exister à l'heure actuelle.

La manière dont l'agent-usager détermine le codage du texte saisi par l'utilisateur est hors du domaine d'application de ce document.

NOTE — les concepteurs de formulaires et de leurs scripts de support devrait être au courant d'un important problème potentiel : lorsque la valeur implicite d'un champ (l'attribut VALUE) est retournée à la soumission (l'utilisateur n'ayant pas modifié cette valeur), il n'y a aucune garantie que la séquence d'octets sera identique à celle du document source, mais seulement comme un codage valide mais peut-être différent de la même séquence d'éléments de texte. Ceci peut se produire même si le codage du document contenant le formulaire et celui utilisé pour sa soumission sont identiques.

Il peut y avoir des différences quand le codage permet qu'une séquence de caractères soit représentée par plusieurs séquences d'octets, et aussi quand une séquence composite (un caractère de base suivi de caractères combinatoires) peut être représentée soit par une autre séquence composite équivalente, soit par un caractère précomposé. Par exemple, la séquence UCS-2 00EA+0323 (LETTRE MINUSCULE LATINE E ACCENT CIRCONFLEXE + DIACRITIQUE POINT SOUSCRIT) peut devenir 1EC7 (LETTRE MINUSCULE LATINE E ACCENT CIRCONFLEXE ET POINT SOUSCRIT), 0065+0302+0323 (LETTRE MINUSCULE LATINE E + DIACRITIQUE ACCENT CIRCONFLEXE + DIACRITIQUE POINT SOUSCRIT), ou d'autres séquences composites équivalentes.

---

6. Du codage externe des caractères

L'inteprétation correcte d'un document textuel exige que le codage de caractères soit connu. Les serveurs HTTP actuels, toutefois, n'ajoute généralement pas le paramètre charset approprié à l'en-tête Content-Type. Ce mauvais comportement est même encouragé par l'existence persistante de fureteurs qui déclarent ne pas reconnaître le type de contenu quand ils reçoivent un paramètre charset. Les implémenteurs d'agents-usager sont fortement encouragés à rendre leur logiciel tolérant envers ce paramètre, même s'il ne peuvent en tirer parti. L'étiquetage correct est très désirable, mais quelques mesures peuvent être prises pour compenser son absence :

Pour les cas où l'on arrive à un document à partir d'un hyperlien dans un document d'origine, un attribut CHARSET est ajouté à la liste d'attributs des éléments formant de tels liens (A et LINK), plus précisément en l'ajoutant à l'entité linkExtraAttributes. La valeur de CHARSET est un paramètre charset MIME, qu'un agent-usager pourra considérer comme un indice du codage de caractères utilisé dans le document auquel l'hyperlien réfère.

Il est possible d'intégrer dans tout document HTML une indication du codage de caractères, en ajoutant aussitôt que possible dans l'élément HEAD un élément META comme suit :

```
<META HTTP-EQUIV="Content-Type"
 CONTENT="text/html; charset=ISO-2022-JP">
```

Cette méthode n'est pas sûre, mais fonctionnera si le codage est tel que les caractères ASCII ne sont pas ambigus au moins jusqu'à l'élément META. Notons qu'un serveur dispose de meilleurs moyens que ce peu fiable META pour connaître le codage de caractères de ses documents ; on trouvera quelques détails et une proposition dans [NICOL2].

En cas de conflit, on considèrera le paramètre charset reçu de la source d'un document comme prioritaire, suivi d'un élément META comme ci-dessus dans le document et finalement de l'éventuel attribut CHARSET d'un hyperlien menant à ce document.

Quand du texte HTML est transmis directement en UCS-2 ou UCS-4, la question de l'ordre des octets se pose : est-ce que l'octet de poids fort de chaque caractère arrive en premier ou en dernier ? Par souci de netteté, il est recommandé de transmettre les octets UCS-2 ou UCS-4 en ordre décroissant de poids, ce qui correspond à l'ordre établi sur Internet pour les quantités à 2 ou 4 octets, à l'exigence de l'ISO 10646 et à la recommandation Unicode pour la transmission

en série et au RFC 1641. Pour optimiser les chances d'interprétation correcte, il est de plus recommandé que les documents en UCS-2 ou UCS-4 débutent toujours par un caractère ESPACE INSÉCABLE À CHASSE NULLE (FEFF ou 0000FEFF hex.) qui, lorsqu'inversé, devient FFFE ou FFFE0000, un caractère garanti inutilisé. Ainsi, un agent-usager qui reçoit FFFE au début d'un texte saura que les octets doivent être inversés pour le reste.

En plus d'UCS-2 et UCS-4, on peut transmettre des données du JUC selon un format de transformation UCS. UTF-7 [RFC1642] et UTF-8 [UTF-8] ont des propriétés favorables (pas de problème d'ordre des octets, différentes saveurs de compatibilité ASCII) qui les rendent dignes d'intérêt, surtout pour la transmission de texte multilingue. Un autre codage, MNEM [RFC1345], a aussi des propriétés intéressantes et est capable de transmettre tout le JUC. Le format de transformation UTF-1 de l'ISO 10646:1993 (enregistré par l'IANA sous le nom ISO-10646-UTF-1), a été retiré de l'ISO 10646 par l'amendement 4, et ne devrait plus être utilisé.

---

7. Texte public d'HTML

7.1. DTD d'HTML

Cette section contient une DTD pour HTML basée sur la DTD HTML 2.0 du RFC 1866, incorporant les changements pour l'apport de fichier précisé dans le RFC 1867 et les changements découlant de ce document.

```
<!--    html-2.x.dtd

        Définition de type de document pour le Langage de balisage à
hyperliens
        internationalisé (DTD HTML)

           Auteurs: Daniel W. Connolly <connolly@w3.org>
                    François Yergeau <yergeau@alis.com>

-->

<!ENTITY % HTML.Version
        "-//IETF//DTD HTML i18n//EN"

           -- Utilisation typique:

               <!DOCTYPE HTML PUBLIC "-//IETF//DTD HTML i18n//EN">
               <html>
               ...
               </html>
           --
           >

<!--============ Entités de vérification de fonctionnalité ==============--
>

<!ENTITY % HTML.Recommended "IGNORE"
        -- Certaines fonctionnalités de ce langage sont nécessaires
           pour être compatibles avec une large utilisation, mais elles
           pourraient compromettre l'intégrité structurelle du document.
```

```
           Cette entité de vérification de fonctionnalité permet de définir
une
           DTD plus stricte qui élimine ces fonctionnalités.
        -->

<![ %HTML.Recommended [
        <!ENTITY % HTML.Deprecated "IGNORE">
]]>

<!ENTITY % HTML.Deprecated "INCLUDE"
        --
           Certaines fonctionnalités du langage sont nécessaires pour
           des raisons de compatibilités avec des versions antérieures de
           ces spécifications. Malheureusement, leurs implémentations
           varient et sont incompatibles, leur utilisation est
déconseillée.
           Cette entité de vérification de fonctionnalité permet de définir
           une DTD qui élimine ces fonctionnalités.
        -->

<!ENTITY % HTML.Highlighting "INCLUDE"
        --
           Utiliser cette entité de vérification de fonctionnalité
           pour s'assurer qu'un document n'utilise pas de balises
           de « rehaut », qui peuvent être ignorées dans les
           implémentations minimales.
        -->

<!ENTITY % HTML.Forms "INCLUDE"
        --
           Utiliser cette entité de vérification de fonctionnalité
           pour s'assurer qu'un document ne comprend pas de
           formulaires qui peuvent ne pas être mis en œuvre
           dans les implémentations minimales.
        -->

<!--============== Noms importés ==============================-->

<!ENTITY % Content-Type "CDATA"
        -- c'est à dire un type de médium internet (également
           connu sous le nom de type de contenu MIME, selon RFC2045)
        -->

<!ENTITY % HTTP-Method "GET | POST"
        -- tel que défini par les spécifications HTTP, RFC 1945
        -->

<!--========== "Macros" DTD ======================-->

<!ENTITY % heading "H1|H2|H3|H4|H5|H6">

<!ENTITY % list " UL | OL | DIR | MENU " >

<!ENTITY % attrs -- attributs communs aux éléments --
         "LANG  NAME     #IMPLIED -- balise de langue selon RFC 1766 --
          DIR  (ltr|rtl) #IMPLIED -- directionnalité du texte --
          ID ID          #IMPLIED -- identificateur d'élément (du RFC1942)
--
          CLASS NAMES    #IMPLIED -- pour surclasser les éléments (du
RFC1942) --">
```

```
<!ENTITY % just -- un attribut pour la justification du texte --
         "ALIGN  (left|right|center|justify)  #IMPLIED">
         -- par défaut gauche pour paragraphes ltr, droite pour rtl -- >

<!--====== Entités mnémoniques de caractères =============->

<!ENTITY % ISOlat1 PUBLIC
  "ISO 8879-1986//ENTITIES Added Latin 1//EN//HTML">
%ISOlat1;
<!-- Voir appendice A -->

<!-- Entités pour les caractères significatifs de balisage -->
<!ENTITY amp CDATA "&#38;"      -- perluette        -->
<!ENTITY gt  CDATA "&#62;"      -- plus grand que   -->
<!ENTITY lt  CDATA "&#60;"      -- plus petit que   -->
<!ENTITY quot CDATA "&#34;"     -- apostrophe       -->

<!-- Entités pour présentation dépendant de la langue (BIDI + analyse
contextuelle) -->
<!ENTITY zwnj CDATA "&#8204;"      -- anti-liant sans chasse -->
<!ENTITY zwj  CDATA "&#8205;"      -- liant sans chasse -->
<!ENTITY lrm  CDATA "&#8206;"      -- marque gauche-à-droite -->
<!ENTITY rlm  CDATA "&#8207;"      -- marque droite-à-gauche -->


<!--========== Entités paramètres d'accès SGML aux documents (SDA) =====-->

<!ENTITY % SDAFORM  "SDAFORM  CDATA  #FIXED"
          -- correspondance bijective        -->
<!ENTITY % SDARULE  "SDARULE  CDATA  #FIXED"
          -- correspondance dépendant du contexte -->
<!ENTITY % SDAPREF  "SDAPREF  CDATA  #FIXED"
          -- préfixe du texte généré     -->
<!ENTITY % SDASUFF  "SDASUFF  CDATA  #FIXED"
          -- suffixe du texte généré     -->
<!ENTITY % SDASUSP  "SDASUSP  NAME   #FIXED"
          -- suspension du processus de transformation -->


!--========== Balisage de texte ====================-->

<![ %HTML.Highlighting [

<!ENTITY % font " TT | B | I ">

<!ENTITY % phrase "EM | STRONG | CODE | SAMP | KBD | VAR | CITE">

<!ENTITY % text "#PCDATA|A|IMG|BR|%phrase|%font|SPAN|Q|BDO|SUP|SUB">

<!ELEMENT (%font;|%phrase) - - (%text)*>
<!ATTLIST ( TT | CODE | SAMP | KBD | VAR )
        %attrs;
        %SDAFORM; "Lit"
        >

<!ATTLIST ( B | STRONG )
        %attrs;
        %SDAFORM; "B"
        >

<!ATTLIST ( I | EM | CITE )
```

```
        %attrs;
        %SDAFORM; "It"
        >

<!-- <TT>            texte à chasse fixe                   -->
<!-- <B>              texte gras                            -->
<!-- <I>              texte italique                      -->

<!-- <EM>             texte mis en valeur                   -->
<!-- <STRONG>         mise en exergue importante            -->
<!-- <CODE>           morceau de code source                          -->
<!-- <SAMP>           échantillon de texte                  -->
<!-- <KBD>            texte saisi au clavier, (c-à-d tapé par
l'utilisateur) -->
<!-- <VAR>            texte variable ou substituable -->
<!-- <CITE>           nom ou titre d'un ouvrage cité -->

<!ENTITY % pre.content "#PCDATA|A|HR|BR|%font|%phrase|SPAN|BDO">

]]>

<!ENTITY % text "#PCDATA|A|IMG|BR|SPAN|Q|BDO|SUP|SUB">

<!ELEMENT BR    - O EMPTY>
<!ATTLIST BR
          %SDAPREF; "&#RE;"
          >

<!-- <BR>       passage à la ligne suivante      -->

<!ELEMENT SPAN - - (text)*>
<!ATTLIST SPAN
        %attrs;
        %SDAFORM; "other #Attlist"
        >

<!-- <SPAN>             conteneur générique                  -->
<!-- <SPAN DIR=...>     Nouvel enchâssement à contre-courant -->
<!-- <SPAN LANG="...">  Langue du contenu                    -->

<!ELEMENT Q - - (%text)*>
<!ATTLIST Q
        %attrs;
        %SDAPREF; '"'
        %SDASUFF; '"'
        >

<!-- <Q>         brève citation                          -->
<!-- <Q LANG=xx> La langue de la citation est xx         -->
<!-- <Q DIR=...> Nouvel enchâssement à contre-courant    -->

<!ELEMENT BDO - - (%text)+>
<!ATTLIST BDO
        LANG   NAME      #IMPLIED
        DIR    (ltr|rtl) #REQUIRED
        ID     ID        #IMPLIED
        CLASS  NAMES     #IMPLIED
        %SDAPREF "Bidi Override #Attval(DIR): "
        %SDASUFF "End Bidi"
        >
```

```
<!-- <BDO DIR=...>   Surcharge la directionnalité du texte selon DIR -->
<!-- <BDO LANG=...>  Langue du contenu                               -->

<!ELEMENT (SUP|SUB) - - (#PCDATA)>
<!ATTLIST SUP
        %attrs;
        %SDAPREF "Superscript(#content)"
        >
<!ATTLIST SUB
        %attrs;
        %SDAPREF "Subscript(#content)"
        >

<!-- <SUP>      Exposant       -->
<!-- <SUB>      Indice         -->


<!--========== Balisage de lien ======================-->

<!ENTITY % linkType "NAME">

<!ENTITY % linkExtraAttributes
        "REL %linkType #IMPLIED
         REV %linkType #IMPLIED
         URN CDATA #IMPLIED
         TITLE CDATA #IMPLIED
         METHODS NAMES #IMPLIED
         CHARSET NAME #IMPLIED
         ">

<![ %HTML.Recommended [
        <!ENTITY % A.content   "(%text)*"
        -- <H1><a name="xxx">En-tête</a></H1>
                est mieux que
           <a name="xxx"><H1>En-tête</H1></a>
        -->
]]>

<!ENTITY % A.content   "(%heading|%text)*">

<!ELEMENT A     - - %A.content -(A)>
<!ATTLIST A
        %attrs;
        HREF CDATA #IMPLIED
        NAME CDATA #IMPLIED
        %linkExtraAttributes;
        %SDAPREF; "<Anchor: #AttList>"
        >

<!-- <A>               Ancre; source/destination d'un hyperlien       -->
<!-- <A NAME="...">    Nom de l'ancre                                 -->
<!-- <A HREF="...">    Adresse de la destination du lien              -->
<!-- <A URN="...">     Adresse permanente de la destination           -->
<!-- <A REL=...>       Rapport à la destination                       -->
<!-- <A REV=...>       Rapport de la destination à ce document        -->
<!-- <A TITLE="...">   Titre de la destination (recommandé)          -->
<!-- <A METHODS="...">  Opérations sur la destination (recommandé)    -->
<!-- <A CHARSET="...">  Jeu de caractères de la destination(recommandé) -->
<!-- <A LANG="...">    Language of contents btw  and           -->
<!-- <A DIR=...>       Contents is a new counterflow embedding        -->
```

```
<!--========== Images ==========================-->

<!ELEMENT IMG    - O EMPTY>
<!ATTLIST IMG
        %attrs;
        SRC CDATA  #REQUIRED
        ALT CDATA #IMPLIED
        ALIGN (top|middle|bottom) #IMPLIED
        ISMAP (ISMAP) #IMPLIED
        %SDAPREF; ``<Fig><?SDATrans Img: #AttList>#AttVal(Alt)</Fig>"
        >

<!-- <IMG>             Image; icône, glyphe ou illustration     -->
<!-- <IMG SRC="...">   Adresse de l'objet-image                 -->
<!-- <IMG ALT="...">   Remplacement textuel                     -->
<!-- <IMG ALIGN=...>   Position relative au texte               -->
<!-- <IMG LANG=...>    L'image contient du texte en cette langue-->
<!-- <IMG DIR=rtl>     L'image agit comme un enchâssement de
                               d-à-g p/r à l'algorithme BIDI            --
>
<!-- <IMG ISMAP>       Chaque pixel peut être indexé            -->

<!--========== Paragraphes =======================-->

<!ELEMENT P     - O (%text)*>
<!ATTLIST P
        %attrs;
        %just;
        %SDAFORM; "Para"
        >

<!-- <P>             Paragraphe                              -->
<!-- <P LANG="...">  Langue du texte du paragraphe           -->
<!-- <P DIR=...>     Directionalité de base du paragraphe    -->
<!-- <P ALIGN=...>   Alignement du paragraphe (justification) -->

<!--========== En-têtes, titres, sections ===============-->

<!ELEMENT HR    - O EMPTY>
<!ATTLIST HR
        DIR  (ltr|rtl) #IMPLIED
        %just;
        %SDAPREF; "&#RE;&#RE;"
        >

<!-- <HR>       filet horizontal -->

<!ELEMENT ( %heading )  - -  (%text;)*>
<!ATTLIST H1
        %attrs;
        %just;
        %SDAFORM; "H1"
        >
<!ATTLIST H2
        %attrs;
        %just;
        %SDAFORM; "H2"
        >
<!ATTLIST H3
        %attrs;
        %just;
```

```
        %SDAFORM; "H3"
        >
<!ATTLIST H4
        %attrs;
        %just;
        %SDAFORM; "H4"
        >
<!ATTLIST H5
        %attrs;
        %just;
        %SDAFORM; "H5"
        >
<!ATTLIST H6
        %attrs;
        %just;
        %SDAFORM; "H6"
        >

<!-- <H1>      En-tête de niveau 1  -->
<!-- <H2>      En-tête de niveau 2  -->
<!-- <H3>      En-tête de niveau 3  -->
<!-- <H4>      En-tête de niveau 4  -->
<!-- <H5>      En-tête de niveau 5  -->
<!-- <H6>      En-tête de niveau 6  -->


<!--========== Flux de texte ======================-->

<![ %HTML.Forms [
        <!ENTITY % block.forms "BLOCKQUOTE | FORM | ISINDEX">
]]>

<!ENTITY % block.forms "BLOCKQUOTE">

<![ %HTML.Deprecated [
        <!ENTITY % preformatted "PRE | XMP | LISTING">
]]>

<!ENTITY % preformatted "PRE">

<!ENTITY % block "P | %list | DL
        | %preformatted
        | %block.forms">

<!ENTITY % flow "(%text|%block)*">

<!ENTITY % pre.content "#PCDATA | A | HR | BR | SPAN | BDO">
<!ELEMENT PRE - - (%pre.content)*>
<!ATTLIST PRE
        %attrs;
        WIDTH NUMBER #implied
        %SDAFORM; "Lit"
        >

<!-- <PRE>              texte pré-formaté                          -->
<!-- <PRE WIDTH=...>    nombre maximal de caractères par ligne     -->
<!-- <PRE DIR=...>      Direction de base du bloc pré-formaté      -->
<!-- <PRE LANG=...>     Langue du contenu                          -->

<![ %HTML.Deprecated [
```

```
<!ENTITY % literal "CDATA"
        --
           historique, mode de syntalisage non-conforme
           où le seul signal de balisage est la balise de
           fin en entier
        -->

<!ELEMENT (XMP|LISTING) - -  %literal>
<!ATTLIST XMP
        %attrs;
        %SDAFORM; "Lit"
        %SDAPREF; "Example:&#RE;"
        >
<!ATTLIST LISTING
        %attrs;
        %SDAFORM; "Lit"
        %SDAPREF; "Listing:&#RE;"
        >

<!-- <XMP>              section modèle           -->
<!-- <LISTING  >        listage informatique     -->

<!ELEMENT PLAINTEXT - O %literal>
<!ATTLIST PLAINTEXT
        %attrs;
        %SDAFORM; "Lit"
        >

<!-- <PLAINTEXT>        passage de texte simple      -->
]]>


<!--========== Listes ==================-->

<!ELEMENT DL    - -  (DT | DD)+>
<!ATTLIST DL
        %attrs;
        COMPACT (COMPACT) #IMPLIED
        %SDAFORM; "List"
        %SDAPREF; "Definition List:"
        >

<!ELEMENT DT    - O (%text)*>
<!ATTLIST DT
        %attrs;
        %SDAFORM; "Term"
        >

<!ELEMENT DD    - O %flow>
<!ATTLIST DD
        %attrs;
        %SDAFORM; "LItem"
        >

<!-- <DL>          Glossaire, ou liste de définition    -->
<!-- <DL COMPACT>  Liste compacte                       -->
<!-- <DT>          terme du glossaire                   -->
<!-- <DD>          définition du terme                  -->

<!ELEMENT (OL|UL) - -  (LI)+>
<!ATTLIST OL
```

```
        %attrs;
        %just;
        COMPACT (COMPACT) #IMPLIED
        %SDAFORM; "List"
        >
<!ATTLIST UL
        %attrs;
        %just;
        COMPACT (COMPACT) #IMPLIED
        %SDAFORM; "List"
        >

<!-- <UL>          Liste non-ordonnée             -->
<!-- <UL COMPACT>  Liste compacte                 -->
<!-- <OL>          Liste ordonnée ou numérotée    -->
<!-- <OL COMPACT>  Liste compacte                 -->

<!ELEMENT (DIR|MENU) - -  (LI)+ -(%block)>
<!ATTLIST DIR
        %attrs;
        %just;
        COMPACT (COMPACT) #IMPLIED
        %SDAFORM; "List"
        %SDAPREF; "<LHead>Directory</LHead>"
        >
<!ATTLIST MENU
        %attrs;
        %just;
        COMPACT (COMPACT) #IMPLIED
        %SDAFORM; "List"
        %SDAPREF; "<LHead>Menu</LHead>"
        >

<!-- <DIR>             Répertoire                     -->
<!-- <DIR COMPACT>     Répertoire sous forme compacte-->
<!-- <MENU>            Menu                           -->
<!-- <MENU COMPACT>    Menu sous forme compacte       -->

<!ELEMENT LI    - O %flow>
<!ATTLIST LI
        %attrs;
        %just;
        %SDAFORM; "LItem"
        >

<!-- <LI>            article de la liste           -->

<!--========== Corps du document  ==================-->

<![ %HTML.Recommended [
        <!ENTITY % body.content "(%heading|%block|HR|ADDRESS|IMG)*"
        -- <h1>En-tête</h1>
           <p>Texte ...
                est mieux que
           <h1>En-tête</h1>
           Texte ...
        -->
]]>

<!ENTITY % body.content "(%heading | %text | %block | HR | ADDRESS)*">
```

```
<!ELEMENT BODY O O  %body.content>
<!ATTLIST BODY
        %attrs;
        >

<!-- <BODY>          Corps du document                  -->
<!-- <BODY DIR=...>  Direction de base de tout le corps -->
<!-- <BODY LANG=...> Langue du contenu                  -->

<!ELEMENT BLOCKQUOTE - - %body.content>
<!ATTLIST BLOCKQUOTE
        %attrs;
        %just;
        %SDAFORM; "BQ"
        >

<!-- <BLOCKQUOTE>       passage cité  -->

<!ELEMENT ADDRESS - - (%text|P)*>
<!ATTLIST  ADDRESS
        %attrs;
        %just;
        %SDAFORM; "Lit"
        %SDAPREF; "Address:&#RE;"
        >

<!-- <ADDRESS>  Adresse, signature  -->


<!--======= Formulaires ====================-->

<![ %HTML.Forms [

<!ELEMENT FORM - - %body.content -(FORM) +(INPUT|SELECT|TEXTAREA)>
<!ATTLIST FORM
        %attrs;
        ACTION CDATA #IMPLIED
        METHOD (%HTTP-Method) GET
        ENCTYPE %Content-Type; "application/x-www-form-urlencoded"
        %SDAPREF; "<Para>Form:</Para>"
        %SDASUFF; "<Para>Form End.</Para>"
        >

<!-- <FORM>                Formulaire à remplir                       -->
<!-- <FORM ACTION="...">   Adresse de renvoi du formulaire            -->
<!-- <FORM METHOD=...>     Méthode de soumission du formulaire        -->
<!-- <FORM ENCTYPE="...">  Représentation des données du formulaire   -->
<!-- <FORM DIR=...>        Direction de base du formulaire            -->
<!-- <FORM LANG=...>       Langue du contenu                          -->

<!ENTITY % InputType "(TEXT | PASSWORD | CHECKBOX |
                       RADIO | SUBMIT | RESET |
                       IMAGE | HIDDEN | FILE )">
<!ELEMENT INPUT - O EMPTY>
<!ATTLIST INPUT
        %attrs;
        TYPE %InputType TEXT
        NAME CDATA #IMPLIED
        VALUE CDATA #IMPLIED
        SRC CDATA #IMPLIED
        CHECKED (CHECKED) #IMPLIED
```

```
        SIZE CDATA #IMPLIED
        MAXLENGTH NUMBER #IMPLIED
        ALIGN (top|middle|bottom) #IMPLIED
        ACCEPT CDATA #IMPLIED --liste de types de contenu --
        ACCEPT-CHARSET CDATA #IMPLIED --liste de jeu de car. acceptés par
serveur --
        %SDAPREF; "Input: "
        >

<!-- <INPUT>                  Champ en entrée du formulaire
-->
<!-- <INPUT TYPE=...>         Type d'interaction de saisie
-->
<!-- <INPUT NAME=...>         Nom du champ d'entrée du formulaire
-->
<!-- <INPUT VALUE="...">      Valeur initiale, sélectionnée ou par défaut
-->
<!-- <INPUT SRC="...">        Adresse de l'image
-->
<!-- <INPUT CHECKED>          État initial est « coché »
-->
<!-- <INPUT SIZE=...>         Indication de la taille du champ
-->
<!-- <INPUT MAXLENGTH=...>    Taille maximale des données
-->
<!-- <INPUT ALIGN=...>        Alignement de l'image
-->
<!-- <INPUT ACCEPT-CHARSET="..."> Liste de jeu de caractères acceptables
-->

<!ELEMENT SELECT - - (OPTION+) -(INPUT|SELECT|TEXTAREA)>
<!ATTLIST SELECT
        %attrs;
        NAME CDATA #REQUIRED
        SIZE NUMBER #IMPLIED
        MULTIPLE (MULTIPLE) #IMPLIED
        %SDAFORM; "List"
        %SDAPREF;
        "<LHead>Select #AttVal(Multiple)</LHead>"
        >

<!-- <SELECT>                 Sélection parmi une liste d'options     -->
<!-- <SELECT NAME=...>        Nom du champ de formulaire              -->
<!-- <SELECT SIZE=...>        Nombre d'options affichées à la fois    -->
<!-- <SELECT MULTIPLE>        Section multiples permises              -->

<!ELEMENT OPTION - O (#PCDATA)*>
<!ATTLIST OPTION
        %attrs;
        SELECTED (SELECTED) #IMPLIED
        VALUE CDATA #IMPLIED
        %SDAFORM; ``LItem''
        %SDAPREF;
        "Option: #AttVal(Value) #AttVal(Selected)"
        >

<!-- <OPTION>                 Une option de la liste de sélection     -->
<!-- <OPTION SELECTED>        État initial                            -->
<!-- <OPTION VALUE=''...''>   Valeur du champ associé à cette option  -->

<!ELEMENT TEXTAREA - - (#PCDATA)* -(INPUT|SELECT|TEXTAREA)>
```

```
<!ATTLIST TEXTAREA
        %attrs;
        NAME CDATA #REQUIRED
        ROWS NUMBER #REQUIRED
        COLS NUMBER #REQUIRED
        ACCEPT-CHARSET CDATA #IMPLIED --liste de jeu de car. acceptés par
serveur --
        %SDAFORM; ``Para''
        %SDAPREF; ``Input Text -- #AttVal(Name): ``
        >

<!-- <TEXTAREA>             Une zone de saisie de texte   -->
<!-- <TEXTAREA NAME=...>    Nom du champ du formulaire    -->
<!-- <TEXTAREA ROWS=...>    Hauteur de la zone            -->
<!-- <TEXTAREA COLS=...>    Largeur de la zone            -->

]]>


<!--======= En-tête de document ======================-->

<![ %HTML.Recommended [
        <!ENTITY % head.extra "">
]]>
<!ENTITY % head.extra "& NEXTID?">

<!ENTITY % head.content "TITLE & ISINDEX? & BASE? %head.extra">

<!ELEMENT HEAD O O  (%head.content) +(META|LINK)>
<!ATTLIST HEAD
        %attrs;
        >

<!-- <HEAD>     En-tête de document   -->

<!ELEMENT TITLE - -  (#PCDATA)* -(META|LINK)>
<!ATTLIST TITLE
        %attrs;
        %SDAFORM; "Ti"
        >

<!-- <TITLE>    Titre du document -->

<!ELEMENT LINK - O EMPTY>
<!ATTLIST LINK
        %attrs;
        HREF CDATA #REQUIRED
        %linkExtraAttributes;
        %SDAPREF; "Linked to : #AttVal (TITLE) (URN) (HREF)>"
        >

<!-- <LINK>                Lien à partir de ce document
-->
<!-- <LINK HREF="...">     Adresse de la destination du lien
-->
<!-- <LINK URN="...">      Nom durable de la destination
-->
<!-- <LINK REL=...>        Rapport de ce document à la destination
-->
<!-- <LINK REV=...>        Rapport de la destination à ce document
-->
```

```
<!-- <LINK TITLE="...">    Titre de la destination (recommandé)
-->
<!-- <LINK CHARSET="...">  Jeu de caractère de la destination (recommandé)
-->
<!-- <LINK METHODS="...">  Opérations permises (recommandé
-->

<!ELEMENT ISINDEX - O EMPTY>
<!ATTLIST ISINDEX
        %attrs;
        %SDAPREF;
        "<Para>[Document is indexed/searchable.]</Para>"
        >

<!-- <ISINDEX>          Ce document est indexable          -->

<!ELEMENT BASE - O EMPTY>
<!ATTLIST BASE
        HREF CDATA #REQUIRED     >

<!-- <BASE>             Document du contexte de base     -->
<!-- <BASE HREF="...">  Adresse de ce document           -->

<!ELEMENT NEXTID - O EMPTY>
<!ATTLIST NEXTID
        N CDATA #REQUIRED
        >

<!-- <NEXTID>          ID suivant à utiliser pour le nom du lien
-->
<!-- <NEXTID N=...>    ID suivant à utiliser pour le nom du lien
-->

<!ELEMENT META - O EMPTY>
<!ATTLIST META
        HTTP-EQUIV  NAME    #IMPLIED
        NAME        NAME    #IMPLIED
        CONTENT     CDATA   #REQUIRED
        >

<!-- <META>                  Métainformation générique                -->
<!-- <META HTTP-EQUIV=...>   Nom de l'en-tête dans la réponse HTTP    -->
<!-- <META NAME=...>         Nom de la  métainformation               -->
<!-- <META CONTENT="...">    Information associée                     -->


<!--======= Structure du document ==================-->

<![ %HTML.Deprecated [
        <!ENTITY % html.content "HEAD, BODY, PLAINTEXT?">
]]>
<!ENTITY % html.content ``HEAD, BODY''>

<!ELEMENT HTML O O  (%html.content)>
<!ENTITY % version.attr "VERSION CDATA #FIXED '%HTML.Version;'">

<!ATTLIST HTML
        %attrs;
        %version.attr;
        %SDAFORM; "Book"
        >
```

7.2. Déclaration SGML pour HTML

```
<!SGML  "ISO 8879:1986"
--
        Déclaration SGML pour le language de balisage hypertexte 2.x
        (HTML 2.x = HTML 2.0 + i18n).

--

CHARSET
         BASESET  "ISO Registration Number 177//CHARSET
                   ISO/IEC 10646-1:1993 UCS-4 with
                   implementation level 3//ESC 2/5 2/15 4/6"
         DESCSET  0   9     UNUSED
                  9   2     9
                  11  2     UNUSED
                  13  1     13
                  14  18    UNUSED
                  32  95    32
                  127 1     UNUSED
                  128 32    UNUSED
                  160 2147483486 160
--
    Dans l'ISO 10646, les positions 0000D800 - 0000DFFF,
    utilisées par le codage UTF-16 d'UCS-4, sont
    réservées, de même que les deux dernières positions
    de chaque plan de l'UCS-4, c'est à dire toutes les
    positions de la forme xxxxFFFE ou xxxxFFFF. Ces
    valeurs codées ou les références numériques
    correspondantes ne doivent pas être incluses dans
    un nouveau document HTML, et devraient être ignorées
    lorsque rencontrées lors du traitement d'un document
    HTML.
--

CAPACITY        SGMLREF
                TOTALCAP        150000
                GRPCAP          150000
               ENTCAP         150000

SCOPE    DOCUMENT
SYNTAX
         SHUNCHAR CONTROLS 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
                17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 127

         BASESET  "ISO 646IRV:1991//CHARSET
                   International Reference Version
                   (IRV)//ESC 2/8 4/2"
         DESCSET  0 128 0

         FUNCTION
                  RE            13
                  RS            10
                  SPACE         32
                  TAB SEPCHAR    9
```

```
         NAMING   LCNMSTRT ""
                  UCNMSTRT ""
                  LCNMCHAR ".-"
                  UCNMCHAR ".-"
                  NAMECASE GENERAL YES
                           ENTITY  NO
         DELIM    GENERAL  SGMLREF
                  SHORTREF SGMLREF
         NAMES    SGMLREF
         QUANTITY SGMLREF
                  ATTSPLEN 2100
                  LITLEN   1024
                  NAMELEN  72    -- plutôt arbitraire; provient des
conventions
                                    de fin de ligne Internet --
                  PILEN    1024
                  TAGLVL   100
                  TAGLEN   2100
                  GRPGTCNT 150
                  GRPCNT   64

FEATURES
  MINIMIZE
    DATATAG  NO
    OMITTAG  YES
    RANK     NO
    SHORTTAG YES
  LINK
    SIMPLE   NO
    IMPLICIT NO
    EXPLICIT NO
  OTHER
    CONCUR   NO
    SUBDOC   NO
    FORMAL   YES
  APPINFO    "SDA"  -- application conforme à SGML Document Access --
>
```

7.3. Jeu d'entités Latin 1

Le texte public suivant donne la liste de tous les caractères du jeu d'entités Added Latin 1, avec leurs noms, syntaxe d'utilisation et descriptions. Cette liste dérive du jeu d'entités ISO 8879:1986//ENTITIES Added Latin 1//EN. HTML reprend le jeu d'entités au complet et y ajoute des entités pour tous les caractères de la partie droite de l'ISO 8859-1.

```
<!-- (C) Organisation internationale de normalisation 1986
      Permission de copie sous toute forme est donnée pour usage avec
      des systèmes et applications SGML conformes à l'ISO 8879, à
      condition que cette notice soit reproduite avec chaque copie.
   -->
 <!-- Jeu d'entités de caractères.  Invocation typique :
      <!ENTITY % ISOlat1 PUBLIC
        "ISO 8879-1986//ENTITIES Added Latin 1//EN//HTML">
      %ISOlat1;
   -->
 <!ENTITY nbsp   CDATA " " -- espace insécable -->
 <!ENTITY iexcl  CDATA "&#161;" -- point d'exclamation inversé -->
```

```
 <!ENTITY cent   CDATA "&#162;" -- symbole centime -->
 <!ENTITY pound  CDATA "&#163;" -- symbole livre -->
 <!ENTITY curren CDATA "&#164;" -- symbole monétaire -->
 <!ENTITY yen    CDATA "&#165;" -- symbole yen -->
 <!ENTITY brvbar CDATA "&#166;" -- barre verticale interrompue -->
 <!ENTITY sect   CDATA "&#167;" -- paragraphe (alinéa) -->
 <!ENTITY uml    CDATA "&#168;" -- tréma -->
 <!ENTITY copy   CDATA "&#169;" -- symbole copyright -->
 <!ENTITY ordf   CDATA "&#170;" -- indicateur ordinal féminin -->
 <!ENTITY laquo  CDATA "&#171;" -- guillemet gauche -->
 <!ENTITY not    CDATA "&#172;" -- signe négation -->
 <!ENTITY shy    CDATA "&#173;" -- trait d'union virtuel -->
 <!ENTITY reg    CDATA "&#174;" -- symbole marque déposée -->
 <!ENTITY macr   CDATA "&#175;" -- macron -->
 <!ENTITY deg    CDATA "&#176;" -- symbole degré -->
 <!ENTITY plusmn CDATA "&#177;" -- signe plus-ou-moins -->
 <!ENTITY sup2   CDATA "&#178;" -- exposant deux -->
 <!ENTITY sup3   CDATA "&#179;" -- exposant trois -->
 <!ENTITY acute  CDATA "&#180;" -- accent aigu -->
 <!ENTITY micro  CDATA "&#181;" -- symbole micro -->
 <!ENTITY para   CDATA "&#182;" -- pied de mouche (fin de paragraphe,
symbole alinéa) -->
 <!ENTITY middot CDATA "&#183;" -- point médian -->
 <!ENTITY cedil  CDATA "&#184;" -- cédille-->
 <!ENTITY sup1   CDATA "&#185;" -- exposant un -->
 <!ENTITY ordm   CDATA "&#186;" -- indicateur ordinal masculin -->
 <!ENTITY raquo  CDATA "&#187;" -- guillemet droit -->
 <!ENTITY frac14 CDATA "&#188;" -- fraction ordinaire un quart -->
 <!ENTITY frac12 CDATA "&#189;" -- fraction ordinaire un demi -->
 <!ENTITY frac34 CDATA "&#190;" -- fraction ordinaire trois quarts -->
 <!ENTITY iquest CDATA "&#191;" -- point d'interrogation inversé -->
 <!ENTITY Agrave CDATA "&#192;" -- A majuscule accent grave -->
 <!ENTITY Aacute CDATA "&#193;" -- A majuscule accent aigu-->
 <!ENTITY Acirc  CDATA "&#194;" -- A majuscule accent circonflexe -->
 <!ENTITY Atilde CDATA "&#195;" -- A majuscule tilde -->
 <!ENTITY Auml   CDATA "&#196;" -- A majuscule tréma -->
 <!ENTITY Aring  CDATA "&#197;" -- A majuscule rond en chef -->
 <!ENTITY AElig  CDATA "&#198;" -- AE  majuscule (ligature) -->
 <!ENTITY Ccedil CDATA "&#199;" -- C majuscule cédille -->
 <!ENTITY Egrave CDATA "&#200;" -- E majuscule accent grave -->
 <!ENTITY Eacute CDATA "&#201;" -- E majuscule accent aigu -->
 <!ENTITY Ecirc  CDATA "&#202;" -- E majuscule accent circonflexe -->
 <!ENTITY Euml   CDATA "&#203;" -- E majuscule tréma -->
 <!ENTITY Igrave CDATA "&#204;" -- I majuscule accent grave -->
 <!ENTITY Iacute CDATA "&#205;" -- I majuscule accent aigu -->
 <!ENTITY Icirc  CDATA "&#206;" -- I majuscule accent circonflexe -->
 <!ENTITY Iuml   CDATA "&#207;" -- I majuscule tréma -->
 <!ENTITY ETH    CDATA "&#208;" -- Eth majuscule (Islandais) -->
 <!ENTITY Ntilde CDATA "&#209;" -- N majuscule tilde -->
 <!ENTITY Ograve CDATA "&#210;" -- O majuscule accent grave -->
 <!ENTITY Oacute CDATA "&#211;" -- O majuscule accent aigu -->
 <!ENTITY Ocirc  CDATA "&#212;" -- O majuscule accent circonflexe -->
 <!ENTITY Otilde CDATA "&#213;" -- O majuscule tilde -->
 <!ENTITY Ouml   CDATA "&#214;" -- O majuscule tréma -->
 <!ENTITY times  CDATA "&#215;" -- signe multiplication -->
 <!ENTITY Oslash CDATA "&#216;" -- O majuscule barré -->
 <!ENTITY Ugrave CDATA "&#217;" -- U majuscule accent grave -->
 <!ENTITY Uacute CDATA "&#218;" -- U majuscule accent aigu -->
 <!ENTITY Ucirc  CDATA "&#219;" -- U majuscule accent circonflexe -->
 <!ENTITY Uuml   CDATA "&#220;" -- U majuscule tréma -->
 <!ENTITY Yacute CDATA "&#221;" -- Y majuscule accent aigu -->
```

```
<!ENTITY THORN  CDATA "&#222;" -- Thorn majuscule (Islandais) -->
<!ENTITY szlig  CDATA "&#223;" -- s dur minuscule (Szet allemand) -->
<!ENTITY agrave CDATA "&#224;" -- a minuscule accent grave -->
<!ENTITY aacute CDATA "&#225;" -- a minuscule accent aigu -->
<!ENTITY acirc  CDATA "&#226;" -- a minuscule accent circonflexe -->
<!ENTITY atilde CDATA "&#227;" -- a minuscule tilde -->
<!ENTITY auml   CDATA "&#228;" -- a minuscule tréma -->
<!ENTITY aring  CDATA "&#229;" -- a minuscule rond en chef -->
<!ENTITY aelig  CDATA "&#230;" -- ae minuscule (ligature) -->
<!ENTITY ccedil CDATA "&#231;" -- c minuscule cédille -->
<!ENTITY egrave CDATA "&#232;" -- e minuscule accent grave -->
<!ENTITY eacute CDATA "&#233;" -- e minuscule accent aigu -->
<!ENTITY ecirc  CDATA "&#234;" -- e minuscule accent circonflexe -->
<!ENTITY euml   CDATA "&#235;" -- e minuscule tréma -->
<!ENTITY igrave CDATA "&#236;" -- i minuscule accent grave -->
<!ENTITY iacute CDATA "&#237;" -- i minuscule accent aigu -->
<!ENTITY icirc  CDATA "&#238;" -- i minuscule accent circonflexe -->
<!ENTITY iuml   CDATA "&#239;" -- i minuscule tréma -->
<!ENTITY eth    CDATA "&#240;" -- eth minuscule (Islandais) -->
<!ENTITY ntilde CDATA "&#241;" -- n minuscule tilde -->
<!ENTITY ograve CDATA "&#242;" -- o minuscule accent grave -->
<!ENTITY oacute CDATA "&#243;" -- o minuscule accent aigu -->
<!ENTITY ocirc  CDATA "&#244;" -- o minuscule accent circonflexe -->
<!ENTITY otilde CDATA "&#245;" -- o minuscule tilde -->
<!ENTITY ouml   CDATA "&#246;" -- o minuscule tréma -->
<!ENTITY divide CDATA "&#247;" -- signe division -->
<!ENTITY oslash CDATA "&#248;" -- o minuscule barré -->
<!ENTITY ugrave CDATA "&#249;" -- u minuscule accent grave -->
<!ENTITY uacute CDATA "&#250;" -- u minuscule accent aigu -->
<!ENTITY ucirc  CDATA "&#251;" -- u minuscule accent circonflexe -->
<!ENTITY uuml   CDATA "&#252;" -- u minuscule tréma -->
<!ENTITY yacute CDATA "&#253;" -- y minuscule accent aigu -->
<!ENTITY thorn  CDATA "&#254;" -- thorn minuscule (Islandais) -->
<!ENTITY yuml   CDATA "&#255;" -- y minuscule tréma -->
```

---

8. Préoccupations de sécurité

Les ancres, les images enchâssées et tous les autres éléments contenant des URI comme paramètres peuvent déclencher le déréférencement de l'URI sur action de l'utilisateur. Dans ce cas, les préoccupations de sécurité du [RFC1738] s'appliquent.

Les méthodes répandues de soumettre des formulaires — HTTP et SMTP — offrent peu de garanties de confidentialité. Les fournisseurs d'information qui demandent de l'information sensible via un formulaire — tout spécialement au moyen d'un champ de saisie de type 'PASSWORD' (cf. la section 8.1.2 du [RFC1866]) — devraient être conscients du manque de confidentialité et en informer leurs utilisateurs.

Bibliographie

[BRYAN88]
M. Bryan, SGML — An Author's Guide to the Standard Generalized Markup Language, Addison-Wesley, Reading, 1988.

[ERCS]

Extended Reference Concrete Syntax for SGML. <http://www.sgmlopen.org/sgml/docs/ercs/ercs-home.html>

[GOLD90]
C. F. Goldfarb, The SGML Handbook, Y. Rubinsky, Ed., Oxford University Press, 1990.

[HTTP-1.1]
R.T. Fielding, H. Frystyk Nielsen, and T. Berners-Lee, Hypertext Transfer Protocol — HTTP/1.1, RFC 2068, Janvier 1997.

[ISO-639]
ISO 639:1988. Norme internationale — Codes pour la représentation des noms de langue. Contenu technique in <http://www.sil.org/sgml/iso639a.html>

[ISO-8859-1]
ISO 8859. Norme internationale — Traitement de l'information — Jeux de caractères graphiques codés sur un seul octet à 8 élément — Partie 1 : Alphabet latin No. 1 (1987) — Partie 2 : Alphabet latin No. 2 (1987) — Partie 3 : Alphabet latin No. 3 (1988) — Partie 4 : Alphabet latin No. 4 (1988) — Partie 5 : Alphabet latin/cyrillique (1988) — Partie 6 : Alphabet latin/arabe (1987) — Partie 7 : Alphabet latin/grec (1987) — Partie 8 : Alphabet latin/hébreu (1988) — Partie 9 : Alphabet latin No. 5 (1989) — Partie 10 : Alphabet latin No. 6 (1992)

[ISO-8879]
ISO 8879:1986. Norme internationale — Traitement de l'information — Texte et systèmes de bureautique — Langage normalisé de balisage généralisé (SGML).

[ISO-10646]
ISO/IEC 10646-1:1993. Norme internationale — Technologie de l'information — Jeu universel de caractères codés sur plusieurs octets (JUC) — Partie 1 : Architecture et plan multilingue de base.

[NICOL]
G.T. Nicol, The Multilingual World Wide Web, Electronic Book Technologies, 1995, <http://www.ebt.com/docs/multling.html>

[NICOL2]
G.T. Nicol, MIME Header Supplemented File Type, Travail en cours, EBT, October 1995.

[RFC1345]
K. Simonsen, Character Mnemonics & Character Sets, RFC 1345, Rationel Almen Planlaegning, June 1992.

[RFC1468]
J. Murai, M. Crispin and E. van der Poel, Japanese Character Encoding for Internet Messages, RFC 1468, Keio University, Panda Programming, June 1993.

[RFC2045]
Freed, N., and N. Borenstein, Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies, RFC 2045, Innosoft, First Virtual, November 1996.

[RFC1641]
D. Goldsmith, M.Davis, Using Unicode with MIME, RFC 1641, Taligent inc., July 1994.

[RFC1642]
D. Goldsmith, M. Davis, UTF-7: A Mail-safe Transformation Format of Unicode, RFC 1642, Taligent inc., July 1994.

[RFC1738]

T. Berners-Lee, L. Masinter, and M. McCahill, Uniform Resource Locators (URL), RFC 1738, CERN, Xerox PARC, University of Minnesota, October 1994.

[RFC1766]
H. Alverstrand, Tags for the Identification of Languages, RFC 1766, UNINETT, March 1995.

[RFC1866]
T. Berners-Lee and D. Connolly, Hypertext Markup Language - 2.0, RFC 1866, MIT/W3C, November 1995.

[RFC1867]
E. Nebel and L. Masinter, Form-based File Upload in HTML, RFC 1867, Xerox Corporation, November 1995.

[RFC1942]
D. Raggett, HTML Tables, RFC 1942, W3C, May 1996.

[RFC2068]
Fielding, R., Gettys, J., Mogul, J., Frystyk, H., and T. Berners-Lee, Hypertext Transfer Protocol — HTTP/1.1, RFC 2068, January 1997.

[SQ91]
SoftQuad, The SGML Primer, 3rd ed., SoftQuad Inc., 1991.

[TAKADA]
Toshihiro Takada, Multilingual Information Exchange through the World-Wide Web, Computer Networks and ISDN Systems, Vol. 27, No. 2, Nov. 1994 , p. 235-241.

[TEI]
TEI Guidelines for Electronic Text Encoding and Interchange. <http://etext.virgina.edu/TEI.html>

[UNICODE]
The Unicode Consortium, The Unicode Standard — Worldwide Character Encoding — Version 1.0, Addison-Wesley, Volume 1, 1991, Volume 2, 1992, and Technical Report #4, 1993. L'algorithme BIDI est dans l'appendice A du volume 1, avec des corrections dans l'appendice D du volume 2.

[UTF-8]
ISO/IEC 10646-1:1993 AMENDMENT 2 (1996). UCS Transformation Format 8 (UTF-8).

[VANH90]
E. van Hervijnen, Practical SGML, Kluwer Academicq Publishers Group, Norwell and Dordrecht, 1990.

Adresses des auteurs

François Yergeau
Alis Technologies
100, boul. Alexis-Nihon, bureau 600
Montréal QC H4M 2P2
Canada

Tél : +1 (514) 747-2547
Fax : +1 (514) 747-2561
Courriel : fyergeau@alis.com

Gavin Thomas Nicol
Electronic Book Technologies, Japan
1-29-9 Tsurumaki,
Setagaya-ku,
Tokyo
Japan

Tél : +81-3-3230-8161
Fax : +81-3-3230-8163
Courriel : gtn@ebt.com, gtn@twics.co.jp

Glenn Adams
Spyglass
118 Magazine Street
Cambridge, MA 02139
U.S.A.

Tél : +1 (617) 864-5524
Fax : +1 (617) 864-4965
Courriel : glenn@spyglass.com

Martin J. Dürst
Multimedia-Laboratory
Department of Computer Science
University of Zurich
Winterthurerstrasse 190
CH-8057 Zurich
Switzerland

Tél : +41 1 257 43 16
Fax : +41 1 363 00 35
Courriel : mduerst@ifi.unizh.ch