Groupe de travail Réseau
Request for Comments : 2277
BCP : 18
Catégorie : Bonnes practiques actuelles

H. Alvestrand
UNINETT
janvier 1998
Traduction Claude Brière de L'Isle

Politique de l'IETF sur les jeux de caractères et les langages

1. Introduction

L'Internet est international.

De ce que l'Internet est international découle une exigence absolue d'échanger des données dans une multitude de langages, qui à leur tour utilisent un nombre incroyable de caractères.

Le présent document décrit les politiques actuelles qui sont appliquées par le groupe de pilotage de l'ingénierie de l'Internet (IESG, *Internet Engineering Steering Group*) envers les efforts de normalisation de l'équipe d'ingénierie de l'Internent (IETF, *Internet Engineering Task Force*) afin d'aider les protocoles de l'Internet à satisfaire à ces exigences.

Le document s'appuie étroitement sur les recommandations de l'atelier sur les jeux de caractères de l'IAB du 29 février au 1er mars 1996, qui est documenté dans la [RFC2130]. Le présent document tente d'être concis, explicite et clair ; ceux qui veulent en savoir plus sont invités à lire la RFC2130.

Le document utilise les termes 'DOIT', DEVRAIT' et 'PEUT' et leurs négations, de la façon décrite dans la [RFC2119]. Dans ce cas, 'la spécification' telle qu'utilisée par la RFC2119 se réfère au traitement des protocoles soumis au processus de normalisation de l'IETF.

2. Où faire l'internationalisation

L'internationalisation est pour les personnes. Cela signifie que les protocoles ne sont pas soumis à l'internationalisation ; les chaînes de texte le sont. Lorsque les éléments de protocole ressemblent à des jetons de texte, comme dans de nombreux protocoles de couche application de l'IETF, les protocoles DOIVENT spécifier quelles parties sont du protocole et quelles parties sont du texte (voir la [RFC2130] § 2.2.1.1).

Les noms posent un problème, parce que les gens y pensent très fortement, que beaucoup d'entre eux sont principalement pour l'usage local, et que tous tendent à se rattacher à tout moment au contexte local. La [RFC1958] recommande l'US-ASCII pour tous les noms à visibilité mondiale.

Le présent document ne rend pas obligatoire une politique de l'internationalisation des noms, mais exige que tous les protocoles décrivent si les noms sont internationalisés ou en US-ASCII.

Note : Dans la pile de protocoles pour toute application, il y a habituellement une ou plusieurs couches qui doivent traiter ces problèmes.

Il ne serait, par exemple, pas approprié de définir des étiquettes de langage pour des trames Ethernet. Mais il est de la responsabilité des groupes de travail de s'assurer que chaque fois que la responsabilité de l'internationalisation est laissée à "une autre couche", ceux qui sont responsables de cette couche sont en fait au courant qu'ils ONT cette responsabilité.

3. Définition des termes

Le présent document utilise le terme de "jeu de caractères" pour signifier un ensemble de règles pour transposer d'une séquence d'octets en une séquence de caractères, comme la combinaison d'un ensemble de caractères codés et d'un schéma de codage de caractères ; c'est aussi ce qui est utilisé comme identifiant dans les paramètres MIME "charset=", et enregistré dans le registre des jeux de caractères de l'IANA [RFC2278]. (Noter que ceci N'EST PAS un terme utilisé par les autres organes de normalisation comme l'ISO).
Pour une définition du terme "jeu de caractères codé", se reporter au rapport du groupe de travail.

Un "nom" est un identifiant comme le nom d'une personne, un nom d'hôte, un nom de domaine, un nom de fichier ou une adresse de messagerie électronique ; il est souvent traité comme un identifiant plutôt que comme du texte, et est souvent utilisé dans les protocoles comme un identifiant pour des entités, sans texte environnant.

3.1 Quel jeu de caractères utiliser

Tous les protocoles DOIVENT identifier, pour toutes les données de caractères, quel jeu de caractères est utilisé.

Les protocoles DOIVENT être capables d'utiliser, pour tout texte, le jeu de caractères UTF-8, qui consiste en le jeu de caractères ISO 10646 codé combiné avec le schéma de codage de caractères UTF-8, comme défini dans [10646] Annexe R (publié dans l'amendement 2).

Les protocoles PEUVENT spécifier, en plus, comment utiliser d'autres jeux de caractères ou d'autres schémas de codage de caractères pour ISO 10646, comme UTF-16, mais l'absence de capacité à utiliser UTF-8 est une violation de cette politique ; une telle violation nécessiterait une procédure de variance ([RFC2026] section 9) avec une justification claire et solide dans le document de spécification du protocole avant d'être entré ou avancé dans le processus de normalisation.

Pour les protocoles existants ou les protocoles qui déplacent des données de mémoires de données existantes, la prise en charge d'autres jeux de caractères, ou même l'utilisation par défaut d'un autre jeu de caractères que UTF-8, peut être une exigence. Ceci est acceptable, mais la prise en charge de UTF-8 DOIT être possible.

Quand ou utilise d'autres jeux de caractères que UTF-8, ceux-ci DOIVENT être enregistrés dans le registre des jeux de caractères de l'IANA, si nécessaire en les enregistrant lors de la publication du protocole.

Note : La norme ISO 10646 appelle le CES UTF-8 un "format de transformation" plutôt qu'un "schéma de codage de caractères", mais cela rentre dans la définition d'un schéma de codage de caractères du rapport de l'atelier de travail).

3.2 Comment décider d'un jeu de caractères

Lorsque le protocole permet un choix parmi plusieurs jeux de caractères, on peut prendre une décision quant au jeu de caractères à utiliser.

Dans certains cas, comme HTTP, il y a une communication directe ou semi directe entre le producteur et le consommateur des données contenant du texte. Dans de tels cas, il peut y avoir du sens à négocier un jeu de caractères avant d'envoyer les données.

Dans d'autres cas, comme la messagerie électronique ou des données mémorisées, il n'y a pas une telle communication, et le mieux qu'on puisse faire est de s'assurer que le jeu de caractères est clairement identifié avec les données mémorisées et de choisir un jeu de caractères qui soit aussi largement connu que possible.

Noter qu'un jeu de caractères est un absolu ; le texte qui est codé dans un jeu de caractères ne peut pas être rendu compréhensible sans la prise en charge de ce jeu de caractères.

(Ceci s'applique aussi aux textes en anglais ; des jeux de caractères tels que l'EBCDIC N'ONT PAS l'ASCII comme sous-ensemble propre.)

Négocier un jeu de caractères peut être vu comme un mécanisme intermédiaire qui doit être pris en charge jusqu'à ce que la prise en charge de l'échange en UTF-8 soit prévalente ; cependant, le cadre temporel de "intermédiaire" peut être d'au moins 50 ans, de sorte qu'il y a toute raison de penser que, en pratique, c'est permanent.

4. Langages

4.1 Besoin d'informations sur le langage

Tous texte lisible par l'homme a un langage.

De nombreuses opérations, incluant un formatage de bonne qualité, la synthèse de texte en parole, la recherche, la mise en forme, la vérification de l'orthographe et ainsi de suite, bénéficient grandement de l'accès à des informations sur le langage d'un morceau de texte, ([RFC2130] § 3.1.1.4).

Les humains ont une certaine tolérance pour les langues étrangères, mais ils sont généralement très mal à l'aise quand on leur présente un texte dans une langue qu'ils ne comprennent pas ; c'est pourquoi la négociation du langage est nécessaire.

Dans la plupart des cas, les machines ne seront pas capables de déduire le langage d'un texte transmis par elles-mêmes ; le protocole doit spécifier comment transférer les informations de langage si il en est de disponibles.

L'interaction entre langage et traitement est complexe ; par exemple, si on compare "name-of-thing(lang=en)" à "name-of-thing(lang=no)" pour égalité, on va généralement s'attendre à une correspondance, alors que le mot "ask(no)" est une espèce d'arbre, et est difficilement utilisable comme verbe de commande.

4.2 Exigence d'étiquetage des langues

Les protocoles qui transfèrent du texte DOIVENT fournir le transport des informations sur le langage de ce texte.

Les protocoles DEVRAIENT aussi assurer le transport des informations sur le langage des noms, lorsque c'est approprié.

Noter que ceci NE signifie PAS que de telles informations doivent toujours être présentes ; l'exigence est que si l'envoyeur des informations souhaite envoyer des informations sur le langage d'un texte, le protocole fournisse un moyen bien défini de transporter ces informations.

4.3 Comment identifier une langue

La [RFC1766] sur les étiquettes de langage est pour le moment l'outil le plus souple disponible pour identifier un langage ; les protocoles DEVRAIENT l'utiliser, ou fournir de claires et solides justifications pour faire autrement dans le document.

Noter aussi qu'un langage est distinct d'une règle locale POSIX ; une règle locale POSIX identifie un ensemble de conventions culturelles, qui peuvent impliquer un langage (la règle locale POSIX ou "C" ne l'implique bien sûr pas) tandis qu'une étiquette de langue décrite dans la RFC1766 identifie seulement un langage.

4.4 Considérations pour la négociation du langage

Les protocoles où les utilisateurs ont du texte qui leur est présenté en réponse à des actions de l'utilisateur DOIVENT fournir la prise en charge de plusieurs langages.

La façon de le faire va varier selon les protocoles ; par exemple, dans certains cas, une négociation où le client propose un ensemble de langages et où le serveur réplique avec l'un d'eux est approprié ; dans d'autres cas, un serveur peut choisir d'envoyer plusieurs variantes d'un texte et laisser le client en prendre une pour l'afficher.

La négociation est utile dans le cas où un côté de l'échange de protocole est capable de présenter le texte en plusieurs langages à l'autre côté, et où l'autre côté a une préférence pour une de celles-ci ; l'exemple le plus courant est le texte qui fait partie des réponses d'erreur, ou les pages de la Toile qui sont disponibles en plusieurs langages.

La négociation d'un langage devrait être considérée comme une exigence permanente du protocole, qui ne disparaîtra jamais à l'avenir.

Dans de nombreux cas, il devrait être possible de l'inclure au titre de l'établissement de la connexion, avec l'authentification et la négociation des autres préférences.

4.5 Langage par défaut

Lorsque du texte lisible par l'homme doit être présenté dans un contexte où l'envoyeur n'a aucune connaissance des préférences de langue du receveur (comme des défaillances de connexion ou des avertissements de messagerie électronique, ou avant la négociation de langage) le texte DEVRAIT être présenté dans le langage par défaut.

Au langage par défaut est allouée l'étiquette "i-default" conformément aux procédures de la RFC1766. Ce n'est pas un langage spécifique, mais elle identifie plutôt une condition où les préférences de langage de l'utilisateur ne peuvent pas être établies.

Les messages en langage par défaut DOIVENT être compréhensibles par un locuteur anglais, car l'anglais est le langage dans lequel, dans le monde entier, le plus grand nombre de personnes seront capables d'obtenir une aide adéquate à interpréter lorsque ils travaillent sur des ordinateurs.

Noter que négocier l'anglais N'EST PAS la même chose que le langage par défaut ; le langage par défaut est une mesure d'urgence dans des situations autrement ingérables .

Dans de nombreux cas, utiliser seulement le texte anglais est raisonnable ; dans certains cas, le texte anglais peut être augmenté de texte dans d'autres langues.

5. Règles locale

La norme [POSIX] définit un concept appelé une "règle locale", qui comporte une quantité d'informations sur l'ordre de collationnement pour le tri, le format de la date, le format de la devise ; et ainsi de suite.

Dans certains cas, et en particulier avec du texte où l'utilisateur est supposé faire un traitement sur le texte, les informations locales peuvent être utilement rattachées au texte ; cela pourrait identifier l'opinion de l'envoyeur sur les règles appropriées à suivre lors du traitement du document, que le receveur peut choisir d'accepter ou d'ignorer.

Le présent document n'exige pas la communication des informations de règles locales sur tout texte, mais encourage leur inclusion lorsque approprié.

Noter que les informations de langage et de jeu de caractères vont souvent être présentes au titre de l'étiquette de règle locale (comme no_NO.iso-8859-1 ; le langage est avant le souligné et le jeu de caractères est après le point) ; il faut veiller à définir précisément quelle spécification de jeu de caractères et de langage s'applique à tout élément de texte.

La règle locale par défaut est la règle locale "POSIX".

6. Documentation des décisions d'internationalisation

Dans les documents qui traitent des questions d'internationalisation, un synopsis des approches choisies pour l'internationalisation DEVRAIT être collecté dans une section appelée "Considérations d'internationalisation ", et placée à côté de la section Considérations pour la sécurité.

Cela donne une référence facile pour ceux qui cherchent un avis sur ces questions lors de la mise en œuvre du protocole.

7. Considérations pour la sécurité

En dehors du fait que des avertissements concernant la sécurité dans une langue étrangère peuvent causer des comportements inappropriés de la part de l'utilisateur, et du fait que les systèmes multilingues ont généralement des problèmes de cohérence entre les variantes de langage, aucune considération de sécurité pertinente n'a été identifiée.

8. Références

[10646] ISO/CEI, “Technologies de l’information – Jeu de caractères universel codé sur plusieurs octets (UCS) - Partie 1 : Architecture et plan de base multilingue", mai 1993, avec amendements.

[POSIX] ISO/CEI 9945-2:1993 “Technologies de l’information – Interface portable de système d’exploitation (POSIX) -- Partie 2 : Enveloppe et utilitaires".

[RFC1766] H. Alvestrand, "Étiquettes pour l'identification des langues", mars 1995. *(Obsolète, voir* RFC5646)

[RFC1958] B. Carpenter, éd., "Principes de l'architecture de l'Internet", juin 1996. (*MàJ par* RFC3439) *(Information)*

[RFC2026] S. Bradner, "Le processus de normalisation de l'Internet -- Révision 3", (BCP0009) octobre 1996. (*MàJ par* RFC3667, RFC3668, RFC3932, RFC3979, RFC3978, RFC5378, RFC6410)

[RFC2119] S. Bradner, "Mots clés à utiliser dans les RFC pour indiquer les niveaux d'exigence", BCP 14, mars 1997.

[RFC2130] C. Weider, C. Preston, K. Simonsen, H. Alvestrand, R. Atkinson, M. Crispin, P. Svanberg, "Rapport de l'atelier Jeux de caractères de l'IAB tenu du 29 février au 1er mars 1996", avril 1997. *(Information)*

[RFC2278] N. Freed, J. Postel, "Procédures d'enregistrement des jeux de caractères par l'IANA", janvier 1998. (*Obsolète, voir* RFC2978) *(Information)*

[RFC2279] F. Yergeau, "UTF-8, un format de transformation de la norme ISO 10646", janvier 1998. (*Obsolète, voir* RFC3629) *(D.S.)*

9. Adresse de l’auteur

Harald Tveit Alvestrand
UNINETT
P.O.Box 6883 Elgeseter
N-7002 TRONDHEIM
NORWAY
téléphone : +47 73 59 70 94
mél : Harald.T.Alvestrand@uninett.no

10. Déclaration complète de droits de reproduction

Remerciement
Le financement de la fonction d'éditeur des RFC est actuellement assuré par la Internet Society.