Internet Engineering Task Force (IETF)
**Request for Comments : 5890**
RFC rendue obsolète : 3490
Catégorie : En cours de normalisation
ISSN: 2070-1721

J. Klensin
août 2010

Traduction Claude Brière de L'Isle

# Noms de domaines internationalisés pour les applications (IDNA) : définitions et cadre documentaire

**Résumé**
Le présent document fait partie d'une collection qui décrit le protocole et le contexte d'usage pour une révision des noms de domaines internationalisés pour les applications (IDNA, *Internationalized Domain Names for Applications*) se substituant à la version antérieure. Il décrit la collection des documents et fournit les definitions et autres matériaux qui sont communs à l'ensemble.

**Statut de ce mémoire**
Ceci est un document de l'Internet en cours de normalisation.

Le présent document a été produit par l'équipe d'ingénierie de l'Internet (IETF). Il représente le consensus de la communauté de l'IETF. Il a subi une révision publique et sa publication a été approuvée par le groupe de pilotage de l'ingénierie de l'Internet (IESG). Tous les documents approuvés par l'IESG ne sont pas candidats à devenir une norme de l'Internet ; voir la Section 2 de la RFC5741.

Les informations sur le statut actuel du présent document, tout errata, et comment fournir des réactions sur lui peuvent être obtenues à http://www.rfc-editor.org/info/rfc5890



## Table des matières

1. Introduction....2
1.1 IDNA2008....2
1.2 Road Map of IDNA2008 Documents....2
2. Définitions et terminologie....3
2.1. Caractères et jeux de caractères....3
2.2. Terminologie relative au DNS....3
3. Considérations relatives à l'IANA....8
4. Considérations sur la sécurité....8
4.1 Questions générales....8
4.2 Longueur des étiquettes U....8
4.3 Problèmes du jeu de caractères local....9

4.4 Caractères visuellement similaires....9
4.5 Recherche IDNA, enregistrement et spécifications de base du DNS....9
4.6 Chaînes d'étiquettes IDN traditionnelles....9
4.7. Différences de la sécurité avec IDNA2003....10
4.8 Résumé....10
5. Remerciements....10
6. Références....10
6.1 Références normatives....10
6.2. Références pour information....11
Adresse de l'auteur....13

# 1. Introduction

## 1.1 IDNA2008

Le présent document fait partie d'une collection qui décrit le protocole et le contexte d'usage d'une révision des noms de domaine internationalisés pour les applications (IDNA, *Internationalized Domain Names for Applications*) qui a été largement achevée en 2008, connue dans la série et ailleurs comme "IDNA2008". La série remplace une version antérieure de IDNA [RFC3490], [RFC3491]. Par convention, cette version de IDNA est appelée dans ces documents "IDNA2003". La plus récente version continue d'utiliser l'algorithme Punycode [RFC3492] et le préfixe de codage comptatible ASCII (ACE, *ASCII-compatible encoding*) provenant de cette version antérieure. La collection des documents est décrite au paragraphe 1.2. Comme il y est indiqué, le présent document donne les définitions et les autres matériaux communs à l'ensemble.

### 1.1.1 Audience

Bien que de nombreuses spécifications de l'IETF soient exclusivement destinées aux développeurs de protocoles, le caractère de IDNA exige qu'il soit compris et correctement utilisé par ceux qui ont la responsabilité des décisions sur :

- quels noms sont permis dans les fichiers de zone du DNS,
- les politiques en matière de noms et de dénominations, et
- le traitement des chaînes de nom de domaine dans les fichiers et systèmes, même sans intention immédiate de les rechercher.

Le présent document et les documents concernés par la définition du protocole, les règles de traitement des chaînes qui incluent des caractères écrits de droite à gauche, et la liste actuelle de caractères et des catégories seront du plus grand intérêt pour les développeurs de protocoles. Le présent document et celui qui contient le matériel explicatif sera de premier intérêt pour les autres, bien qu'ils puissent avoir à se reporter aux autres documents de l'ensemble pour certains détails qui sont en référence.

Le présent document et ceux qui lui sont associés sont écrits dans la perspective d'un utilisateur qui est au courant de IDNA, application, ou mise en œuvre. Bien qu'il puisse réitérer des règles et exigences fondamentales du DNS pour l'agrément du lecteur, il ne tente pas d'être exhaustif sur les principes du DNS et ne devrait pas être considéré comme substitut d'une compréhension en profondeur des protocoles et spécifications du DNS.

### 1.1.2 Langage normatif

Les mots clés "DOIT", "NE DOIT PAS", "EXIGE", "DEVRA", "NE DEVRA PAS", "DEVRAIT", "NE DEVRAIT PAS", "RECOMMANDE", "PEUT", et "FACULTATIF" en majuscules dans ce document sont à interpréter comme décrit dans le BCP 14, [RFC2119].

## 1.2 Panorama des documents IDNA2008

IDNA2008 consiste en les documents suivants :

- Le présent document, contient les définitions et autres matériaux qui sont nécessaires pour comprendre les autres documents de l'ensemble. Dans les autres documents, on s'y réfère de façon informelle comme "Définitions".

- Un document, la [RFC5894], qui donne une vue d'ensemble du protocole et des tableaux associés avec le matériel explicatif et des explications des décisions qui ont conduit à IDNA2008. Ce document contient aussi des avis sur le fonctionnement des registres et pour ceux qui utilisent les noms de domaines internationalisés (IDN, *Internationalized*

*Domain Names*). On s'y réfère de façon informelle dans les autres documents de l'ensemble comme "Raisons". Il n'est pas normatif.

- Un document, la [RFC5891], qui décrit le cœur du protocole IDNA2008 et son fonctionnement. En combinaison avec le document Bidi, décrit immédiatement ci-dessous, il met explicitement à jour et remplace la [RFC3490]. On s'y réfère de façon informelle dans les autres documents de l'ensemble comme "le protocole".

- Un document, la [RFC5893], qui spécifie les règles particulières (Bidi) pour les étiquetttes qui contiennent des caractères qui sont écrits de droite à gauche.

- Une spécification, la [RFC5892], des catégories et règles qui identifient les codets permis dans une étiquette écrite dans la forme des caractères d'origine (définis plus spécifiquement comme "étiquette U" au paragraphe 2.3.2.1) sur la base des allocations de codets Unicode 5.2 [Unicode52] et des règles supplémentaires uniques de IDNA2008. Les règles fondées sur Unicode sont supposées être stables à travers les mises à jour de Unicode et donc indépendantes des versions de Unicode. Cette spécification rend obsolète la [RFC3941] et l'utilisation par IDN des tableaux auxquels il se réfère. On s'y réfère de façon informelle dans les autres documents de l'ensemble comme "Tableaux".

- Un document, la [RFC5895] qui expose le problème de la transposition des caractères en d'autres caractères et qui donne des lignes directrices pour le faire quand c'est approprié. Ce document, auquel on se réfère informellement comme "Transposition", donne des avis ; il n' est pas une partie exigée de IDNA.

## 2. Définitions et terminologie

### 2.1. Caractères et jeux de caractères

Un codet est une valeur d'entier dans l'espace de code d'un jeu de caractères codé. Dans Unicode, ce sont des entiers de 0 à 0x10FFFF.

Unicode version 5.2 [Unicode52] est un jeu de caractères codé contenant un peu plus de 100 000 caractères auxquels sont alloués des codets. Un seul codet Unicode est noté dans les présents documents par "U+" suivi par quatre à six chiffres hexadécimaux, tandis qu'une gamme de codets Unicode est notée par des nombres hexadécimaux de quatre à six chiffres séparés par "..", sans préfixe.

ASCII signifie US-ASCII [ASCII], un jeu de caractères codé contenant 128 caractères associés aux codets dans la gamme 0000..007F. Unicode est un surensemble de l'ASCII et peut être vu comme en étant une généralisation ; il inclut tous les caractères ASCII et les associe aux codets équivalents.

Les "lettres" sont, de façon informelle, des généralisations de l'ASCII et de la compréhension au sens courant de ce terme, c'est-à-dire les caractères qui sont utilisés pour écrire du texte et ne sont pas des chiffres, des symboles, ou de la ponctuation. Formellement, ce sont des caractères d'une valeur de catégorie générale Unicode commençant à "L" (voir le paragraphe 4.5 de la norme Unicode [Unicode52]).

### 2.2. Terminologie relative au DNS

Quand il discute du DNS, le présent document suppose généralement la terminologie utilisée dans les spécifications du DNS [RFC1034] [RFC1035] comme modifiées par les [RFC1123] et [RFC2181]. Le terme de "recherche" est utilisé pour décrire la combinaison des opérations effectuées par le protocole IDNA2008 et celles effectuées en fait par un résolveur DNS. Le processus de placer une entrée dans le DNS est appelée un "enregistrement". C'est similaire à l'usage contemporain courant de ce terme dans d'autres contextes. par conséquent, toute administration de zone du DNS est décrite comme un "registre", et les termes de "registre" et "administrateur de zone" sont utilisés de manière interchangeable, sans considération des arrangements administratifs réels ou du niveau dans l'arborescence du DNS. Plus de détails sur ces relations figurent dans le document Raisons.

Le terme de "codet LDH" est défini dans le présent document comme se référant aux codets associés aux lettres ASCII (codets Unicode 0041..005A et 0061..007A), chiffres (0030..0039), et le caractère tiret - moins (U+002D). "LDH" est une abréviation pour "lettres, chiffres, tiret" mais est utilisé spécifiquement dans le présent document pour se référer à l'ensemble des règles de dénomination décrites au paragraphe 2.3.1.

Les spécifications de base du DNS, [RFC1034] et [RFC1035], discutent des "noms de domaines" et des "noms d'hôte", mais de nombreuses personnes utilisent les termes de façon interchangeable, comme le font des sections de ces

spécifications. Le manque de clarté sur cette terminologie a contribué à la confusion sur les intentions dans certains cas. Ces documents utilisent généralement le terme "nom de domaine". Quand ils se réfèrent, par exemple, à des restrictions de la syntaxe des noms d'hôtes, ils citent explicitement les documents de définition pertinents. Le reste des définitions de ce paragraphe est essentiellement une révison : si il y a des différences entre ces définitions et les définitions des documents de base du DNS ou de ceux cités ci-dessous, les définitions des autres documents ont la préséance.

Une étiquette *(label)* est un composant individuel d'un nom de domaine. Les étiquettes sont généralement montrées séparées par des points ; par exemple, le nom de domaine "www.exemple.com" est composé de trois étiquettes : "www", "exemple", et "com". (La convention complète de nom qui utilise un point final décrite dans la [RFC1123], qui peut être explicite comme dans "www.exemple.com." ou implicite comme dans "www.exemple.com", n'est pas prise en considération dans la présente spécification.) IDNA étend l'ensemble des caractères utilisables dans les étiquettes qui sont traitées comme du texte (à la différence des étiquettes de chaîne binaire discutées dans les [RFC1035] et [RFC2181] et chaînes de bits [RFC2673]) mais seulement dans certains contextes. Les différents contextes pour les différents ensembles de caractères utilisables sont précisés dans le paragraphe suivant. Pour le reste du présent document et dans ce qui s'y rapporte, le terme "étiquette" est l'abréviation de "étiquette de texte", et "chaque étiquette" signifie "chaque étiquette de texte", incluant le contexte étendu.

### 2.3 Terminologie spécifique de IDNA

**Cett**e section définit une terminologie pour réduire la dépendance aux termes et définitions qui ont posé des problèmes dans le passé. La relation entre ces définitions est illustrée dans les figures 1 et 2. Dans la première de ces figures, les numéros entre parenthèses se réfèrent aux notes en dessous de la figure.

#### 2.3.1 Étiquettes LDH

C'est la forme d'étiquette classique utilisée, bien qu'avec quelques restrictions supplémentaires, sur les noms d'hôtes [RFC0952]. Sa syntaxe est identique à celle décrite dans la "syntaxe de nom préférée " du paragraphe 3.5 de la [RFC1034] telle que modifiée par la [RFC1123]. En bref, c'est une chaîne consistant en lettres, chiffres et tiret ASCII, avec la restriction supplémentaire que le tiret ne peut pas apparaître au début ou à la fin de la chaîne. Comme toutes les étiquettes du DNS, sa longueur totale ne doit pas excéder 63 octets.

Les étiquettes LDH incluent les étiquettes spécialisées utilisées par IDNA (décrites comme des "étiquettes A" ci dessous) et certaines formes interdites supplémentaires (aussi décrites ci-dessous).

Pour faciliter la clarté de la description, deux nouveaux sous ensembles d'étiquettes LDH sont créés par l'introduction de IDNA. Elles sont appelées étiquettes LDH réservées (étiquettes LDH-R) et étiquettes LDH non réservées (étiquettes LDH-NR). Les étiquettes LDH réservées, appelées des "noms de domaines étiquetés" dans certains autres contextes, ont comme propriété de contenir "--" dans leurs troisième et quatrième caractères mais qui par ailleurs se conforment aux règles des étiquettes LDH. Seul un sous ensemble des étiquettes LDH-R peut être utilisé dans les applications à capacité IDNA. Ce sous ensemble consiste en la classe d'étiquettes qui commencent par le préfixe "xn--" (indépendant de la casse) mais qui autrement se conforment aux règles des étiquettes LDH. Ce sous ensemble est appelé "étiquettes XN" dans cet ensemble de documents. Les étiquettes XN sont encore subdivisées en celles dont les caractères restants (après le "xn--") sont des sorties valides de l'algorithme Punycode [RFC3492] et celles dont ils ne le sont pas (voir ci-dessous). Les étiquettes XN qui sont des sorties valides de Punycode sont appelées des "étiquettes A" si elles satisfont aussi aux autres critères pour la validité de IDNA décrites ci-dessous. Parce que les étiquettes LDH (et, bien sûr, toute étiquette DNS) ne doivent pas faire plus de 63 octets, la portion d'une étiquette XN déduite de l'algorithme Punycode est limitée à pas plus de 59 caractères ASCII. Les étiquettes LDH non éservées sont l'ensemble des étiquettes LDH valides qui n'ont pas "--" dans les troisièmes et quatrièmes positions.

Une conséquence des restrictions sur les caractères valides dans le jeu de caractères Unicode natifs (voir les étiquettes U) se trouve être que les annotation de casse mixte, de la sorte mentionnée dans l'Appendice A de la [RFC3492], ne sont jamais utiles. Donc, comme une étiquette A valide est le résultat du codage Punycode d'une étiquette U, les étiquettes A ne devraient être produites qu'en minuscules, en dépit d'une correspondance avec d'autres (de casse mixte ou en majuscules) étiquettes potentielles dans le DNS.

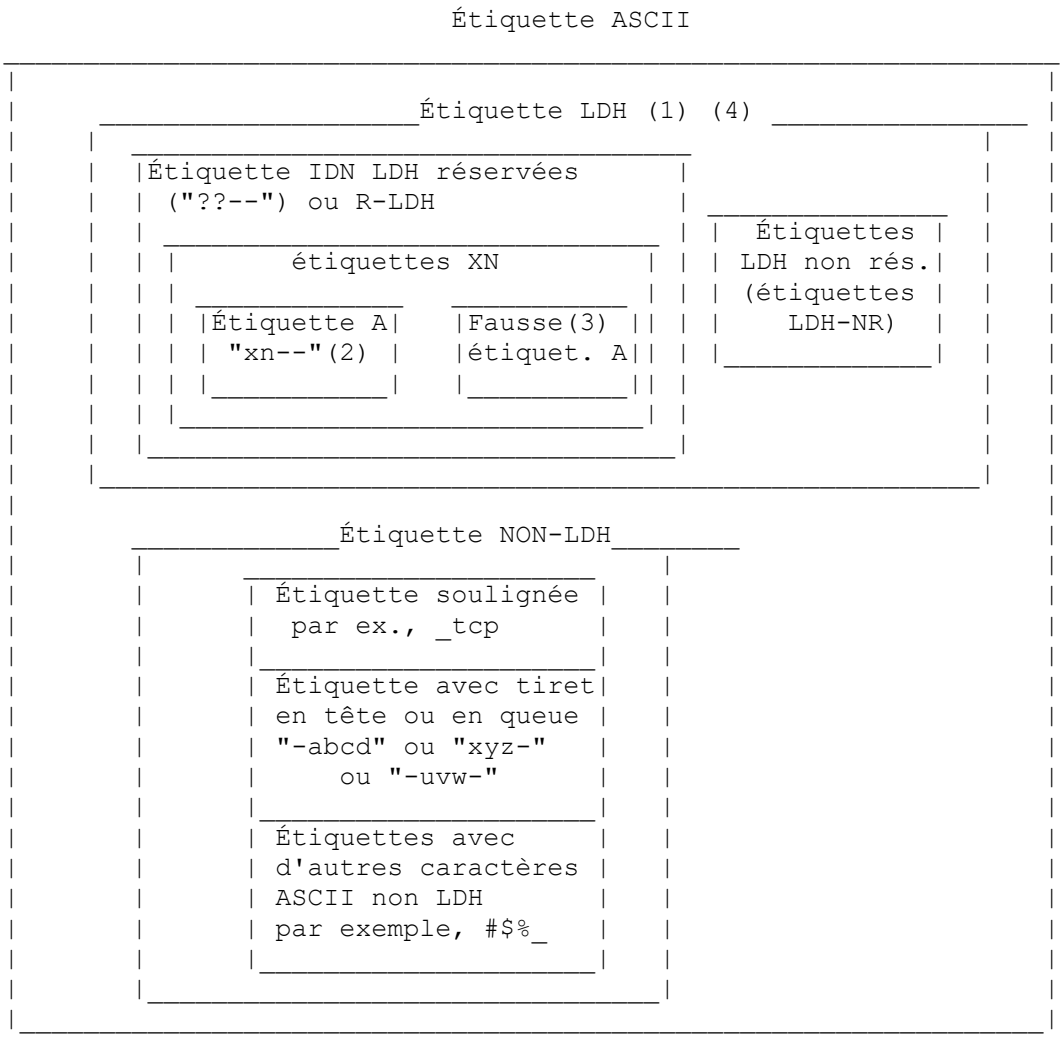
Certaines chaînes qui sont préfixées de "xn--" pour former des étiquettes peuvent n'être pas le résultat de l'algorithme Punycode, peuvent échouer aux autres vérifications mentionnées ci-dessous, ou peuvent violer d'autres restrictions d'IDNA et donc ne sont aussi pas des étiquettes IDNA valides. On les appelle par facilité de "fausses étiquettes A".

Les étiquettes au sein de la classe d'étiquettes LDH R qui ne sont pas préfixées de "xn--" sont aussi des étiquettes IDNA non valides. Pour permettre une future utilisation de mécanismes similaires à IDNA, ces étiquettes NE DOIVENT PAS être

traitées comme des étiquettes LDH ordinaires par les programmes qui se conforment à IDNA et NE DEVRAIENT PAS être mélées à des étiquettes IDNA dans la même zone.

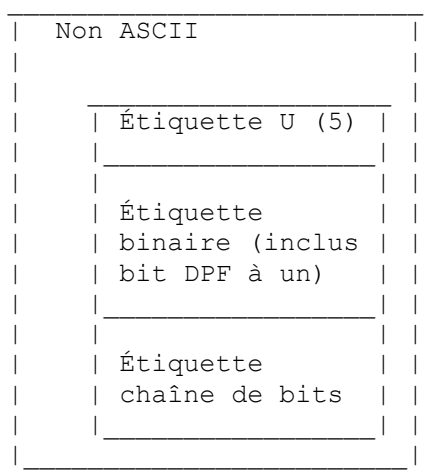
Ces distinctions entre les étiquettes LDH possibles n'ont de signification que pour les logiciels qui sont à capacité IDNA ou pour de futures extensions qui se fondent sur le modèle de même "préfixe et codage". Pour les systèmes à capacité IDNA, les types d'étiquette valides sont : les étiquettes A, les étiquettes U, et les étiquettes LDH-NR.

Les étiquettes IDNA ont deux nuances : une forme codée en ACE et une forme Unicode (caractère natif). On les appelle respectivement les étiquettes A et les étiquettes U, et elles sont décrites en détail au paragraphe suivant.




(1) Lettres ASCII (majuscules et minuscules) chiffres, tirets. Le tiret ne peut pas appaaraître en première ou dernière position. Pas plus de 63 octets.
(2) Noter que la chaîne suivant "xn--" doit être le résultat valide de l'algorithme Punycode et doit être convertible en forme d'étiquette U valide.
(3) Noter qu'une fausse étiquette A a un préfixe "xn--" mais le reste de l'étiquette N'EST PAS le résultat valide de l'algorithme Punycode.
(4) Les sous types d'étiquette LDH ne peuvent pas être distingués par les applications sous capacité IDNA.

**Figure 1 : Espace de terminologie IDNA et DNS -- étiquettes ASCII**




(5) Pour les applications sans capacité IDNA, les étiquettes U sont indistinguables des binaires.

**Figure 2 : Les étiquettes non ASCII**

### 2.3.2 Termes pour les codages d'étiquette IDN

#### 2.3.2.1 Chaînes IDNA valides, étiquettes A et U

Pour les applications à capacité IDNA, les trois types d'étiquettes valides sont "les étiquettes A", "les étiquettes U", et "les étiquettes LDH-NR", dont chacun est défini ci-dessous. Les relations entre elles sont illustrées aux Figure 1 et Figure 2.

- Une chaîne est "à capacité IDNA" si elle satisfait toutes les exigences de ces spécifications d'étiquette IDNA. Les chaînes à capacité IDNA peuvent apparaître dans l'une des deux formes définies ci-dessous, ou peuvent être tirées du sous ensemble d'étiquettes LDH-NR. Les chaînes à capacité IDNA doivent aussi se conformer à toutes les exigences de base du DNS pour les étiquettes. Ces documents font une référence spécifique à la forme appropriée à tout contexte dans lequel la distinction est importante.

- Une "étiquette A" est la forme en codage compatible ASCII (ACE, *ASCII-Compatible Encoding*) (voir le paragraphe 2.3.2.5) d'une chaîne à capacité IDNA. Ce doit être une étiquette complète : IDNA est défini pour des étiquettes, pas pour des morceaux d'étiquettes et pas pour des noms de domaines complets. Cela signifie, par définition, que chaque étiquette A va commencer par le préfixe IDNA ACE, "xn--" (voir le paragraphe 2.3.2.5) suivi par une chaîne qui est un résultat valide de l'algorithme Punycode [RFC3492] et donc d'un maximum de 59 caractères ASCII. Le préfixe et la chaîne doivent ensemble se conformer à toutes les exigences pour une étiquette qui peut être mémorisée dans le DNS incluant la conformité aux règles pour les étiquettes LDH (paragraphe 2.3.1). Si et seulement si une chaîne satisfaisant aux exigences ci-dessus peut être décodée en une étiquette U, c'est une étiquette A.

- Une "étiquette U" est une chaîne IDNA valide de caractères Unicode en forme de normalisation C (NFC, *Normalization Form C*) et incluant au moins un caractère non ASCII, exprimé en forme de codage Unicode standard (comme l'UTF-8). Elle est aussi soumise aux contraintes sur les caractères permis qui sont spécifiés au paragraphe 4.2 du document de protocole et aux règles des Sections 2 et 3 du document Tableaux, aux contraintes de Bidi de ce document si elle contient des caractères provenant de scripts qui sont écrits de droite à gauche, et aux contraintes de symétrie décrites immédiatement ci-dessous. Les conversions entre étiquettes U et étiquettes A sont effectuées selon la spécification "Punycode" [RFC3492], en ajoutant ou supprimant le préfixe ACE en tant que de besoin.

Pour être valides, les étiquettes U et A doivent obéïr à une importante contrainte de symétrie. Bien que cette contrainte puisse être vérifiée de plusieurs façons, une étiquette A A1 doit pouvoir être produite en la convertissant à partir d'une étiquette U U1, et cette étiquette U U1 doit pouvoir être produite par la conversion d'une étiquette A A1. Entre autres choses, cela implique que les deux étiquettes U et A doivent être des chaînes en Unicode NFC [UAX15] en forme normalisée. Ces chaînes DOIVENT contenir seulement des caractères spécifiés ailleurs dans la présente série de documents, et seulement dans les contextes indiqués comme appropriés.

Toutes les règles ou conventions qui s'appliquent aux étiquettes DNS en général s'appliquent à celle des étiquettes U ou A qui serait la plus restrictive. Il y a deux exceptions à ce principe. D'abord, la restriction aux caractères ASCII ne s'applique pas aux étiquette U. Ensuite, l'expansion de la forme d' étiquette A à une étiquette U peut produire des chaînes qui sont beaucoup plus longues que la limite normale de 63 octets du DNS (potentiellement jusqu'à 252 caractères) due à l'efficacité

de la compression de l'algorithme Punycode. De telles étiquettes U de longueur étendue sont valides du point de vue de IDNA, mais il faut faire attention car des limites plus courtes peuvent être imposées par certaines applications.

Pour le contexte, les applications qui n'ont pas de capacité IDNA traitent toutes les étiquettes LDH comme valides pour leur apparition dans les fichiers de zone et les interrogations du DNS et certaines d'entre elles peuvent permettre des types d'étiquettes supplémentaires (c'est-à-dire, ne pas imposer la restriction LDH). Les applications à capacité IDNA permettent seulement aux étiquettes A et LDH-NR d'apparaître dans les fichiers de zone et les interrogations. Les étiquettes U peuvent apparaître, ainsi que les deux autres, en formes de présentation et d'interface d'utilisateur, et dans les protocoles qui utilisent les formes IDNA mais n'impliquent pas le DNS lui-même.

Spécifiquement, pour les applications et contextes à capacité IDNA, les trois catégories admises sont les étiquettes A, les étiquettes U, et les étiquettes LDH-NR. Parmi les étiquettes LDH réservées (étiquettes LDH R) seules les étiquettes A sont valides pour l'utilisation de IDNA.

Les chaînes qui apparaissent comme étant des étiquettes A ou des étiquettes U sont traitées dans diverses opérations du document de protocole [RFC5891]. Il n'est pas encore démontré que ces chaînes soient conformes aux conditions mentionnées ci-dessus parce que elles sont en cours de validation. De telles chaînes peuvent être appelées "non validées", "putatives", ou "apparentes", ou comme étant "dans la forme de" un des types d'étiquette pour indiquer qu'elles n'ont pas encore été vérifiées comme satisfaisant aux exigences de conformité spécifiées.

Les étiquettes A non validées sont seulement connues comme étiquettes XN, tandis que les fausses étiquettes A ont été démontrées échouer à certaines des vérifications d'étiquette A. De façon similaire les étiquettes U non validées sont simplement des étiquettes non ASCII qui peuvent ou non satisfaire aux exigences pour les étiquettes U.

#### 2.3.2.2 Étiquette LDH-NR

Ces spécifications utilisent le terme de "étiquette LDH-NR" de façon stricte pour se référer à une étiquette toute ASCII qui obéit à la syntaxe d'étiquette LDH discutée au paragraphe 2.3.1 et qui n'est ni une IDN ni une forme d'étiquette réservée par IDNA (étiquette LDH-R). On doit souligner que toutes les étiquettes A obéissent aux règles de "nom d'hôte" [RFC0952] autres que la restriction de longueur de ces règles.

#### 2.3.2.3 Nom de domaine internationalisé et étiquette internationalisée

Un "nom de domaine internationalisé" (IDN) est un nom de domaine qui contient au moins une étiquette A ou U, mais qui autrement peut contenir tout mélange d'étiquettes LDH-NR, d'étiquettes A, ou étiquettes U. Tout comme cela a été le cas avec les noms ASCII, certains administrateurs de zone du DNS peuvent imposer des restrictions, au delà de celles imposées par le DNS ou IDNA, sur les caractères ou chaînes qui peuvent être enregistrées comme étiquettes dans leur zones. À cause de la diversité des caractères qui peuvent être utilisés dans une étiquette U et de la confusion qu'elle peut occasionner, de telles restrictions sont obligatoires pour les registres et zones des IDN même si les restrictions particulières ne font pas partie de ces spécifications (la question est discutée plus en détails au paragraphe 4.3 du document de protocole [RFC5891]. Parce que ces restrictions, couramment appelées "restrictions de registre", affectent seulement ce qui peut être enregistré et pas le traitement de recherche, elles n'ont pas d'effet sur la syntaxe ou la sémantique des messages de protocole du DNS ; une interrogation sur un nom qui ne correspond à aucun enregistrement va donner la même réponse sans considération de la raison pour laquelle il n'est pas dans la zone. Les clients qui produisent des interrogations ou qui interprètent les réponses ne peuvent pas être supposés avoir connaissance des restrictions ou conventions spécifiques de la zone. Voir un exposé plus détaillé dans la section sur la politique d'enregistrement dans le document Raisons [RFC5894].

Les "étiquettes internationalisées" sont utilisées quand un terme doit se référer à une seule étiquette d'un IDN, c'est-à-dire, une qui peut être une étiquette LDH-NR, une étiquette A, ou une étiquette U. Il y a des formats d'étiquettes DNS standard, comme les "étiquettes avec soulignement" utilisées pour les enregistrements de localisation de service (SRV) [RFC2782], qui ne rentrent dans aucune de ces trois catégories et donc ne sont pas des étiquettes internationalisées.

#### 2.3.2.4 Équivalence d'étiquette

Dans IDNA, l'équivalence des étiquettes est définie dans les termes des étiquettes A. Si les étiquettes A sont égales dans une comparaison indépendante de la casse, alors les étiquettes sont considérées comme équivalentes, sans se soucier de la façon dont elles sont représentées. À cause de l'isomorphisme des étiquettes A et U dans IDNA2008, il est possible de comparer les étiquettes U directement ; voir les détails dans le document de protocole [RFC5891]. Les étiquettes LDH traditionnelles ont déjà une notion d'équivalence : au sein de cette liste de caractères, majuscules et minuscules sont considérées comme équivalentes. La notion IDNA d'équivalence est une extension de cette notion plus ancienne mais,

parce que le protocole ne spécifie aucune transposition obligatoire et que seulement ces formes isomorphes sont considérées, les seules équivalences sont :

- o correspondance exacte (identité de la chaîne de bits) entre une paire d'étiquettes U,
- o correspondance entre une paire d'étiquettes A, en utilisant les règles de correspondance normales du DNS indépendamment de la casse,
- o équivalence entre une étiquette U et une étiquette A déterminée en traduisant la forme d'étiquette U en une forme d'étiquette A et ensuite en vérifiant la correspondance entre les étiquettes A en utilisant les règles de correspondance normales du DNS indépendamment de la casse.

#### 2.3.2.5 Préfixe ACE

Le "préfixe ACE" est défini dans le présent document comme étant une chaîne de caractères ASCII, "xn--", qui apparaît au début de chaque étiquette A (ACE, *ASCII-Compatible Encoding*, codage compatible ASCII).

#### 2.3.2.6 Créneau de nom de domaine

Un "créneau de nom de domaine" est défini dans le présent document comme un élément de protocole ou un argument de fonction ou une valeur de retour (et ainsi de suite) explicitement conçu pour porter un nom de domaine. Des exemples de créneaux de nom de domaine incluent le champ QNAME d'une interrogation au DNS, l'argument de nom gethostbyname() ou getaddrinfo() des fonctions standard de bibliothèque C, la partie d'une adresse de messagerie électronique qui suit le caractère "@" dans le paramètre SMTP MAIL ou les commandes RCPT ou le champ "From:" d'un en-tête de message électronique, et la portion hôte de l'URI dans l'attribut "src" d'une étiquette HTML "<IMG>". Une chaîne qui a la syntaxe d'un nom de domaine mais apparait en texte général n'est pas dans un créneau de nom de domaine. Par exemple, un nom de domaine qui apparaît dans le corps en clair d'un message électronique n'occupe pas un créneau de noms de domaine.

Un "créneau de nom de domaine à capacité IDNA" est défini pour cet ensemble de documents comme étant un créneau de nom de domaine explicitement conçu pour porter un nom de demaine internationalisé comme défini dans le présent document. La désignation peut être statique (par exemple, dans la spécification du protocole ou de l'interface) ou dynamique (par exemple, en résultat d'une négociation dans une session interactive).

Les créneaux de noms qui ne sont pas à capacité IDNA incluenr évidemment tout créneau de nom de domaine dont la spécification inclut IDNA. Noter que les exigences de certains protocoles qui utilisent le DNS pour la mémorisation de données empêchent l'utilisation des IDN. Par exemple, le format exigé pour les étiquettes à soulignement utilisées par le protocole de localisation de service [RFC2782] interdit la représentation d'une étiquette non ASCII dans le DNS en utilisant des étiquettes A parce que ces étiquettes en relation avec le SRV doivent commencer par un souligné. Bien sûr, les étiquettes IDN non ASCII peuvent faire partie d'un nom de domaine qui inclut aussi des étiquettes à souligné.

### 2.3.3 Ordre des caractères dans les étiquettes

Parce que les étiquettes IDN peuvent contenir des caractères qui sont lus, et de préférence affichés, de droite à gauche, il y a une ambiguité potentielle sur quel caractère est "premier" dans une étiquette. Pour les besoins de ces spécifications, les étiquettes sont examinées, et les caractères numérotés, strictement dans l'ordre dans lequel ils apparaissent "sur le réseau". Cet ordre est équivalent au traitement du caractère le plus à gauche comme le premier d'une étiquette qui est lue de gauche à droite et du caractère le plus à droite comme étant le premier dans une étiquette qui est lue de droite à gauche. La spécification Bidi contient des développement supplémentaires sur les conditions qui influencent l'ordre.de lecture

### 2.3.4 Punycode est un algorithme, pas un nom ni un adjectif

Il y a eu un peu de confusion sur la question de savoir si une "chaîne Punycode" inclut ou non le préfixe ACE et si il est exigé qu'une telle chaîne puisse être le résultat de l'opération ToASCII (voir la Section 4 de la [RFC3490]). La présente spécification déconseille l'utilisation du terme "Punycode" pour décrire quelque chose d'autre que la méthode de codage et l'algorithme de la [RFC3492]. Les termes définis ci-dessus sont préférés comme beaucoup plus clairs que le terme de "chaîne Punycode".

# 3. Considérations relatives à l'IANA

Les actions de l'IANA pour la présente version de IDNA (IDNA2008) sont spécifiées dans le document Tableaux [RFC5892]. Une vue d'ensemble des relations entre les divers registres de l'IANA apparaît dans le document Raisons [RFC5894]. Le présent document ne spécifie aucune action de la part de l'IANA.

# 4. Considérations sur la sécurité

## 4.1 Questions générales

La sécurité de l'Internet repose partiellement sur le DNS. Donc, tout changement des caractéristiques du DNS peut changer la sécurité d'une grande partie de l'Internet.

Les noms de domaines sont utilisés par les utilisateurs pour identifier et se connecter aux hôtes de l'Internet et aux autres ressources du réseau. La sécurité de l'Internet est compromise si un utilisateur qui entre un seul nom internationalisé est connecté à des serveurs différents sur la base d'interprétations différentes du nom de domaine internationalisé. En plus des caractères permis par IDNA2003 et ses conventions de transposition (voir le paragraphe 4.6) la spécification actuelle change l'interprétation de quelques caractères qui étaient transposés en d'autres dans la version antérieure ; les administrateurs de zone devraient être conscients des problèmes que cela peut soulever et prendre les mesures appropriées. Le contexte de ce problème est discuté plus en détails dans le document Raisons [RFC5894].

En plus des matériaux des considérations sur la sécurité qui apparaissent dans le présent document, le document Bidi [RFC5893] contient une discussion des problèmes de sécurité spécifiques des étiquettes qui contiennent des caractères provenant de scripts qui sont normalement écrits de droite à gauche.

## 4.2 Longueur des étiquettes U

Les étiquettes associées au DNS ont traditionellement été limitées à 63 octets par les restrictions générales de la RFC 1035 et par le besoin de les traiter comme des chaînes de six bits suivies par la chaîne dans l'appel réel au DNS. Ce format est utilisé dans quelques autres applications et, en général, ces représentations de noms de domaines comme des étiquettes séparées par des points et comme des paires de longueur-chaîne ont été traitées comme interchangeables. Parce que les étiquettes A (la forme actuellement utilisée dans le DNS) sont potentiellement beaucoup plus compressées que l'UTF-8 (et l'UTF-8 est, en général, plus compressé que l'UTF-16 ou l'UTF-32) les étiquettes U qui obéissent à toutes les contraintes pertinentes de symétrie (et autres) de ces documents peuvent être nettement plus longues, potentiellement jusqu'à 252 caractères (codets Unicode). Un nom de domaine pleinement qualifié contenant plusieurs de ces étiquettes peut évidemment aussi excéder la limite nominale de 255 octets pour ces noms. Les auteurs d'applications qui utilisent des étiquettes U doiventt faire très attention à éviter des débordements de mémoire tampon et des erreurs de troncature et les attaques dans des contextes où des chaînes plus courtes sont attendues.

## 4.3 Problèmes du jeu de caractères local

Quand les systèmes utilisent des jeux de caractères locaux autres que ASCII et Unicode, ces spécifications laissent le problème de la conversion entre le jeu de caractère local et Unicode à l'application ou au système local. Si des applications différentes (ou des versions différentes d'une application) mettent en œuvre des règles différentes pour convertir des jeux de caractères codés, elles pourraient interpréter le même nom de façons différentes et contacter des serveurs différents. Ce problème n'est pas résolu par les protocoles de sécurité, comme la sécurité de la couche Transport (TLS) [RFC5246], qui ne prend pas en compte les jeux de caractères locaux.

## 4.4 Caractères visuellement similaires

Pour aider à empêcher la confusion entre des caractères visuellement similaires (parfois appelés "confusionnels") il est suggéré que les mises en œuvre fournissent des indications visuelles où un nom de domaine contient plusieurs scripts, en particulier lorsque les scripts contiennent des caractères qu'on peut facilement confondre visuellement, comme un omicron grec mélé à du texte latin. De tels mécanismes peuvent aussi être utilisés pour montrer quand un nom contient un mélange de caractères chinois simplifiés avec des caractères traditionnels qui ont des formes simplifiées, ou pour distinguer zéro et un du "O" majuscule et du "L" minuscule. Les administrateurs de zone du DNS peuvent imposer des restrictions (soumises aux limitations identifiées ailleurs dans ces documents) qui essayent de minimiser les caractères qui ont une apparence similaire ou des interprétations similaires.

Si plusieurs caractères apparaissent dans une étiquette et que l'étiquette consiste seulement en caractères dans un seul script, les caractères individuels qui pourraient être confondus avec d'autres si ils étaient comparés séparément peuvent être non ambigus et ne pas préter à confusion. Par ailleurs, cette observation rend les étiquettes qui contiennent des caractères provenant de plus d'un script (souvent appelées "étiquettes de scripts mixtes") encore plus risquées -- les utilisateurs vont tendre à voir ce qu'ils s'attendent à voir et le contexte est un renforcement puissant de la perception. En même temps, alors que les risques associés aux étiquettes de scripts mixtes sont clairs, les interdire simplement ne va pas éliminer les problèmes, en particulier lorsque des scripts en rapports étroits sont utilisés. Par exemple, il y a de nombreuses chaînes entièrement dans des scripts grecs ou cyrilliques qui peuvent être confondues les unes avec les autres ou avec des chaînes de script latin.

Il vaut de noter qu'il n'y a pas de solution technique complète aux problèmes de confusion de caractères. On peut réduire l'extension des problèmes de diverses façons, mais probablement jamais l'éliminer. Certaines suggestions spécifiques sur l'identification et le traitement des caractères confondables apparaît dans une publication du Consortium Unicode [UTR36].

### 4.5 Recherche IDNA, enregistrement et spécifications de base du DNS

La spécification du protocole [RFC5891] décrit les procédures pour enregistrer et rechercher les étiquettes qui ne sont pas compatibles avec la syntaxe préférée décrite dans les spécifications de base du DNS (voir au paragraphe 2.3.1) parce qu'elles contiennent des caractères non ASCII. Ces procédures dépendent de l'utilisation d'une forme de codage spéciale compatible ASCII qui ne contient que des caractères permis dans les noms d'hôtes par ces spécifications antérieures. Le codage utilisé est Punycode [RFC3492]. Aucun problème de sécurité comme des augmentations de longueur de chaîne ou de nouvelles valeurs permises n'est introduit par le processus de codage ou l'utilisation de ces valeurs codées, à part ceux introduites par le codage ACE lui-même.

Les noms de domaines (ou des portions d'eux) sont parfois comparés à un ensemble de domaines auxquels donner un traitement spécial si une correspondance se produit, par exemple, traités comme plus privilégiés que d'autres ou bloqués d'une certaine façon. Dans de telles situations, il est particulièrement important que les comparaisons soient faites correctement, comme spécifié dans la section "Exigences" de la [RFC5891]. Pour les étiquettes qui sont déjà en forme ASCII, la comparaison appropriée se réduit à la même comparaison ASCII insensible à la casse qui a toujours été utilisée pour les étiquettes ASCII bien que les applications à capacité IDNA soient supposées ne rechercher que les étiquettes A et LDH-NR, c'est-à-dire, éviter de rechercher les étiquettes LDH R qui ne sont pas des étiquettes A.

L'introduction de IDNA signifie que toutes les étiquettes existantes qui commencent par le préfixe ACE vont être analysées comme des étiquettes A, au moins jusqu'à une échoue aux essais pertinents, que ce soit ou non l'intention de l'administrateur de zone ou de l'enregistrant. Il n'est pas prouvé que cela ait causé de problèmes pratiques depuis l'adoption de la RFC 3490, mais le risque existe quand même en principe.

### 4.6 Chaînes d'étiquettes IDN traditionnelles

La norme d'URI [RFC3986] et un certain nombre de spécifications d'applications (par exemple, SMTP [RFC5321] et HTTP [RFC2616]) ne permettent pas d'étiquettes non ASCII dans les noms du DNS utilisés avec ces protocoles, c'est-à-dire, seule la forme d'étiquette A des IDN est permise dans ces contextes. Si seulement des étiquettes A sont utilisées, les différences d'interprétation entre IDNA2003 et la présente version ne se manifestent que pour les caractères dont l'interprétation a en fait changé (par exemple, des caractères, comme ZWJ et ZWNJ, qui n'étaient transposés en rien dans IDNA2003 et qui sont considérés comme légitimes dans certains contextes par ces spécifications). En dépit de cette interdiction, il y a un nombre significatif de fichiers et bases de données de l'Internet dans lesquels les chaînes de noms de domaine apparaissent en forme de caractère native ; un sous ensemble de ces chaînes utilise des étiquettes de caractères natifs qui exigent une transposition IDNA2003 pour produire des étiquettes A valides. Le traitement de telles étiquettes va varier selon les types d'applications et les préférences du concepteur d'application : dans certaines situations, des avertissements à l'utilisateur ou le rejet direct peuvent être appropriés ; dans d'autres, il peut être préférable de tenter d'appliquer les transpositions antérieures si une recherche strictement conforme à ces spécifications échoue ou même pour faire des recherches selon les deux ensembles de règles. Cette situation générale est discutée plus en détails dans la [RFC5894]. Cependant, en l'absence de directives des registres sur la façon dont sont traitées les chaînes qui pourraient avoir des interprétations différentes avec IDNA2003 et avec la présente spécification, il est possible que les différences puissent être utilisées comme composante d'une attaque de correspondance de noms ou de confusion de noms. Il est donc approprié d'y faire attention.

### 4.7. Différences de la sécurité avec IDNA2003

Les modèles d'enregistrement et de recherche décrits dans cet ensemble de documents changent les mécanismes disponibles pour les applications de recherche pour déterminer la validité des étiquettes qu'elles rencontrent. À certains égards, la capacité d'essais est renforcée. Par exemple, des étiquettes éventuelles qui contiennent des codets non alloués vont maintenant être rejetées, alors que IDNA2003 les permettait (voir dans la [RFC5894] une discussion des raisons de cela). Par ailleurs, la spécification du protocole ne suppose plus que l'application qui cherche un nom sera capable de déterminer, et appliquer, les informations sur la version du protocole utilisée dans l'enregistrement. En théorie, cela peut augmenter les risques car l'application sera capable de faire moins de validation avant la recherche. En pratique, la protection fournie par cet essai peut être largement illusoire pour les raisons expliquées dans la [RFC4690] et ailleurs dans ces documents.

Tout changement à la procédure Stringprep [RFC3454] qui est profilée et utilisée dans IDNA2003, ou, plus généralement dans le modèle de l'IETF de l'utilisation des chaînes de caractères internationalisées dans les différents protocoles, crée des risques de changements involontaires à ces protocoles, invalidant les applications ou bases de données déployées, et ainsi de suite. Mais ces spécifications ne changent pas du tout Stringprep ; elles le court-circuitent simplement. Parce que ces documents ne dépendent pas de Stringprep, la question de mettre à niveau d'autres protocoles qui ont cette dépendance peut être laissée aux experts de ces protocoles : les changement de IDNA et les possibles mises à niveau des protocoles ou conventions de sécurité sont des questions indépendantes.

### 4.8 Résumé

Aucun mécanisme impliquant des noms ou identifiants ne peut seul protéger contre une large variété de menaces et attaques contre la sécurité qui sont largement indépendantes du système de désignation ou d'identification. Ces attaques incluent des pages contrefaites, la capture d'interrogations au DNS et leur détournement, et ainsi de suite.

## 5. Remerciements

La version initiale du présent document a largement été créée en extrayant du texte des premières versions du projet de document Raisons [RFC5894] ; voir sa section du même nom et celle intitulée "Contributeurs".

Des suggestions textuelles spécifiques après le processus d'extraction ont été faites par Vint Cerf, Lisa Dusseault, Bill McQuillan, Andrew Sullivan, et Ken Whistler. D'autres changements ont été faits en réponse à des commentaires plus généraux, des listes de problèmes ou des erreurs spécifiques provenant des participants au groupe de travail et d'autres observateurs, parmi lesquels Lyman Chapin, James Mitchell, Subramanian Moonesamy, et Dan Winship.

## 6. Références

### 6.1 Références normatives

[ASCII] American National Standards Institute , "USA Code for Information Interchange", ANSI X3.4-1968. ANSI X3.4-1968 a été remplacé par des versions plus récentes avec de légères modifications, mais la version 1968 reste celle de référence pour l'Internet.

[RFC1034] P. Mockapetris, "Noms de domaines - Concepts et facilités", STD 13, novembre 1987. (*MàJ par* RFC1101, 1183, 1348, 1876, 1982, 2065, 2181, 2308, 2535, 4033, 4034, 4035, 4343, 4035, 4592, 5936, 8020, 8482, 8767)

[RFC1035] P. Mockapetris, "Noms de domaines – Mise en œuvre et spécification", STD 13, novembre 1987. (*MàJ par* RFC1101, 1183, 1348, 1876, 1982, 1995, 1996, 2065, 2136, 2181, 2137, 2308, 2535, 2673, 2845, 3425, 3658, 4033, 4034, 4035, 4343, 5936, 5966, 6604, 7766, 8482, 8767

[RFC1123] R. Braden, éditeur, "Exigences pour les hôtes Internet – Application et prise en charge", STD 3, octobre 1989. *(MàJ par* RFC7766)

[RFC2119] S. Bradner, "Mots clés à utiliser dans les RFC pour indiquer les niveaux d'exigence", BCP 14, mars 1997. *(MàJ par* RFC8174)

[UAX15] The Unicode Consortium, "Unicode Standard Annex #15: Unicode Normalization Forms, Revision 31", septembre 2009, < http://www.unicode.org/reports/tr15/tr15-31.html >.

[Unicode52] The Unicode Consortium, "The Unicode Standard, Version 5.2.0", 2009. ISBN 978-1-936213-00-9). <http://www.unicode.org/versions/Unicode5.2.0/>.

### 6.2. Références pour information

[RFC0952] K. Harrenstien, M. Stahl, E. Feinler, "Spécification du tableau des hôtes de l'Internet du DOD", octobre 1985.

[RFC2181] R. Elz et R. Bush, "Clarifications pour la spécification du DNS", juillet 1997. *(P.S., MàJ par* RFC4035, RFC2535, RFC4343, RFC4033, RFC4034, RFC5452, RFC8767)

[RFC2616] R. Fielding et autres, "Protocole de transfert hypertexte -- HTTP/1.1", juin 1999. *(D.S., MàJ par* 2817, 6585)

[RFC2673] M. Crawford, "Étiquettes binaires dans le système des noms de domaine", août 1999. *(Remplacée par* RFC6891)

[RFC2782] A. Gulbrandsen, P. Vixie et L. Esibov, "Enregistrement de ressource DNS pour la spécification de la localisation des services (DNS SRV)", février 2000.

[RFC3454] P. Hoffman et M. Blanchet, "Préparation de chaînes internationalisées ("stringprep")", décembre 2002. *(P.S.)*

[RFC3490] P. Faltstrom et autres, "Internationalisation des noms de domaine dans les applications (IDNA)", mars 2003. (*Remplacée par les* RFC5890 et 5891, *P.S.*)

[RFC3491] P. Hoffman et M. Blanchet, "Nameprep : Profil Stringprep pour les noms de domaine internationalisés (IDN)", mars 2003. (*Remplacée par la* RFC5891, *P.S.*)

[RFC3492] A. Costello, "Punycode : Codage Bootstring d'Unicode pour les noms de domaine internationalisés dans les applications (IDNA)", mars 2003. *(P.S.)*

[RFC3986] T. Berners-Lee, R. Fielding et L. Masinter, "Identifiant de ressource uniforme (URI) : Syntaxe générique", STD 66, janvier 2005.

[RFC4690] J. Klensin et autres, "Révisons et recommandations pour les noms de domaines internationalisés (IDN)", septembre 2006. *(Information)*

[RFC5246] T. Dierks, E. Rescorla, "Version 1.2 du protocole de sécurité de la couche Transport (TLS)", août 2008. *(P.S. ; remplace* RFC3268, 4346, 4366 ; *MàJ* RFC4492 ; *rendue obsolète par la* RFC8446)

[RFC5321] J. Klensin, "Protocole simple de transfert de messagerie", octobre 2008. (*Remplace* RFC2821) (*MàJ* RFC1123) *(D.S.)*

[RFC5891] J. Klensin, "Noms de domaine internationalisés pour les applications (IDNA) : Le protocole", août 2010. (*Remplace* RFC3490, RFC3491) (*MàJ* RFC3492) *(P.S.)*

[RFC5892] P. Faltstrom, "Codets Unicode et noms de domaine internationalisés pour les applications (IDNA)", août 2010. *(P.S. ; MàJ par RFC8753)*

[RFC5893] H. Alvestrand, C. Karp, "Écritures de droite à gauche pour les noms de domaine internationalisés pour les applications (IDNA)", août 2010. *(P.S.)*

[RFC5894] J. Klensin, "Noms de domaine internationalisés pour les applications (IDNA) : Fondements, explication, et motivations", août 2010. *(Information)*

[RFC5895] P. Resnick, P. Hoffman, "Transposition de caractères pour les noms de domaine internationalisés pour les applications (IDNA)", septembre 2010. *(Information)*

[UTR36] The Unicode Consortium, "Unicode Technical Report #36: Unicode Security Considerations, Revision 7", juillet 2008, < http://www.unicode.org/reports/tr36/tr36-7.html >.

## Adresse de l'auteur

John C Klensin
1770 Massachusetts Ave, Ste 322
Cambridge, MA  02140
USA
téléphone : +1 617 245 1457
mél : john+ietf@jck.com