Petit guide de l'acceptation universelle

ACCEPTATION

VALIDATION

STOCKAGE

TRAITEMENT

AFFICHAGE

Les logiciels et les services en ligne prennent en charge l'acceptation universelle lorsqu'ils offrent les fonctionnalités énumérées ci-dessus pour tous les noms de domaine et adresses électroniques.

[1]L'ASCII est le codage de caractères traditionnellement utilisé sur l'Internet, défini dans la norme Internet RFC 20 (https://tools.ietf.org/html/rfc20). L'Unicode est défini par le Consortium Unicode (http://unicode.org).

L'« acceptation universelle », c'est quoi ?

Certains logiciels ne reconnaissent pas ou ne traitent pas correctement l'ensemble des noms de domaine et des adresses électroniques. Les noms de domaine peuvent comprendre des chaînes de premier niveau qui ne sont plus les anciennes chaînes que l'on connaît, et les noms de domaine et les adresses électroniques peuvent désormais utiliser des caractères tirés d'un répertoire basé sur Unicode bien plus grand que l'ASCII traditionnel.[1] L'**acceptation universelle** (UA) correspond à l'état dans lequel tous les noms de domaine et les adresses électroniques valides sont **acceptés, validés, stockés, traités** et **affichés** correctement et de façon cohérente.

Le Groupe directeur sur l'acceptation universelle (UASG) est une initiative communautaire visant à sensibiliser ainsi qu'à identifier et résoudre les problèmes liés à l'acceptation universelle de l'ensemble des noms de domaine et des adresses électroniques. Le but de cette initiative est d'aider à garantir une expérience uniforme et positive pour les internautes à travers le monde. Elle est soutenue par l'ICANN (la Société pour l'attribution des noms de domaine et des numéros sur Internet) et ses participants sont issus de plus de 200 organisations du monde entier, dont Afilias, Apple, CNNIC, GoDaddy, Google, Microsoft et Verisign. Pour de plus amples informations sur l'UASG et ses récentes évolutions, consultez : **www.uasg.tech**.

Ce petit guide décrit les recommandations de l'UASG visant à parvenir à l'acceptation universelle dans le cadre des cinq processus (acceptation, validation, stockage, traitement et affichage) dans lesquels des systèmes sont confrontés à des noms de domaine et des adresses électroniques. Il est destiné aux dirigeants et managers chargés des activités liées aux technologies de l'information et à la conception de logiciels. Il présente les recommandations de l'UASG avec précision sans pour autant donner certains détails qui seraient importants pour un architecte logiciel ou un ingénieur. Vous trouverez ces détails dans le document UASG 007 « Introduction à l'acceptation universelle ».

ACCEPTATION

L'**acceptation** est le processus via lequel une adresse électronique ou un nom de domaine est reçu d'une interface utilisateur, d'un fichier ou d'une API (interface de programmation d'application) utilisé par une application logicielle ou un service en ligne.

Recommandations de l'UASG

- Les champs de saisie doivent être assez grands pour accepter toute entrée valide. Selon le type de codage, un nom de domaine peut exiger pas moins de 670 octets. Une adresse électronique peut avoir une partie locale (la partie avant l'arobase) comprenant jusqu'à 64 octets en plus d'un nom de domaine, pour une longueur totale allant jusqu'à 735 octets.
- Les applications et les services doivent accepter les noms de domaine et adresses électroniques codés par UTF-8[2] et doivent reconnaître que le nombre d'octets occupés par le codage UTF-8 peut être supérieur au nombre de caractères affichés.
- Un IDN peut être saisi et affiché soit dans son script d'origine soit dans une version ASCII conçue dans une optique de rétrocompatibilité, par exemple 测试 et xn--0zwm56d. Le codage Unicode du script d'origine est appelé étiquette U ; le codage compatible avec l'ASCII équivalent est appelé étiquette A.[3] Les logiciels doivent accepter les étiquettes A et les étiquettes U, mais la conversion des étiquettes A en étiquettes U à des fins d'affichage et de traitement ne nécessite pas d'étiquettes A.
- Dans presque tous les cas, un nom de domaine ou une adresse électronique saisi doit être converti en forme de normalisation Unicode C (NFC)[4] avant tout traitement. La NFC n'étant pas sans perte, il peut s'avérer nécessaire dans de rares circonstances de reporter la normalisation jusqu'à ce qu'un traitement ait établi le ou les processus spécifiques dans lesquels elle doit être appliquée.

[2]UTF-8 code chaque point de code Unicode en tant que séquence d'un à quatre octets. Il est défini dans le RFC 3629.

[3]La conversion entre étiquettes A et étiquettes U est effectuée par l'algorithme « Punycode » défini dans le RFC 3492 et le RFC 5891.

[4]Voir l'annexe n° 15 de la norme Unicode, « Formes de normalisation Unicode » (https://www.unicode.org/reports/tr15/tr15-47.html)..

VALIDATION

La **validation** est le processus visant à vérifier la syntaxe d'une adresse électronique ou d'un nom de domaine, et l'existence d'un nom qui est censé exister dans le DNS. Des validations techniques peuvent devoir être mises à jour afin de pouvoir travailler avec des noms de domaine et des adresses électroniques modernes.

[5]Voir les RFC 5890, 5891, 5892, 5893 et 5894 pour la définition de l'IDNA2008.

[6]Voir la « Liste des domaines de premier niveau » (https://www.icann.org/resources/pages/tlds-2012-02-25-en).

Recommandations de l'UASG

- Les entrées doivent être validées de manière adéquate selon l'usage qui doit en être fait. Tous les noms de domaine doivent être validés conformément à la norme relative aux noms de domaine internationalisés dans les applications (actuellement IDNA2008).[5] Cela garantit la validité syntaxique du nom.
- Si une chaîne d'entrée est censée être une entrée existante dans le DNS, validez-la via une recherche dans le DNS.
- Si une chaîne d'entrée est censée être un nom de domaine valide qui pourrait ne pas (encore) se trouver dans le DNS, il peut quand même être possible d'en valider une partie. Par exemple, le nom de domaine de premier niveau (TLD) peut être vérifié par rapport à la liste faisant autorité des noms de TLD valides dressée par l'Autorité chargée de la gestion de l'adressage sur Internet (IANA).[6]
- Afin de valider une adresse électronique, validez la partie du domaine tel que décrit ci-dessus. Dans la mesure où la partie locale d'une adresse électronique est uniquement définie par le système de messagerie qui reçoit les courriers électroniques, il n'est généralement pas possible de la valider. Le fait de demander à l'utilisateur de saisir l'adresse électronique deux fois pourrait permettre de détecter les fautes de frappe.
- Dans la plupart des cas, tous les composants d'un nom de domaine ou d'une adresse électronique (sauf le nom de TLD s'il ne s'agit pas d'un IDN) doivent être en un seul script (par exemple arabe ou han) ou en des scripts étroitement liés (par exemple kanji japonais, katakana, hiragana et romaji). Utilisez la norme technique Unicode n° 39, « Mécanismes de sécurité Unicode » (https://unicode.org/reports/tr39/#Restriction_Level_Detection), afin de vérifier que les scripts d'une séquence Unicode respectent les bonnes pratiques.

STOCKAGE

Le **stockage** désigne le stockage temporaire ou à long terme de noms de domaine et d'adresses électroniques qui doivent être stockés sous des formats bien définis indépendamment de la durée du stockage escomptée.

Recommandations de l'UASG

- Dans presque tous les cas, les noms de domaine ou adresses électroniques doivent être normalisés conformément à la forme de normalisation Unicode C (NFC) avant le stockage. La NFC n'étant pas sans perte, il peut s'avérer nécessaire dans de rares circonstances de reporter la normalisation jusqu'à ce qu'un traitement ait établi le ou les processus spécifiques dans lesquels elle doit être appliquée.
- Dans la plupart des applications, les noms de domaine et adresses électroniques doivent être stockés dans des fichiers et bases de données codés en tant qu'UTF-8, le codage Unicode le plus courant et le mieux pris en charge. Dans certains cas, lorsque le logiciel doit interagir avec des bases de données historiques, il peut être plus simple d'utiliser le même codage que celui de la base de données.
- Dans le code d'application, la représentation la plus adaptée d'Unicode dépend de l'environnement de programmation. Un grand nombre de langages de programmation usuels, dont les langages de script Python et Perl, ont un support intégré pour la conversion Unicode et automatique vers ou depuis UTF-8 sur les entrées et les sorties.
- Les applications doivent choisir une représentation interne cohérente (soit des étiquettes U soit des étiquettes A) pour les IDN. Puisque toute étiquette U peut être transformée en une étiquette A unique et vice versa, les deux formes sont acceptables.

TRAITEMENT

Le **traitement** intervient lorsqu'une adresse électronique ou un nom de domaine est utilisé par une application ou un service afin de mener une activité (par exemple la recherche ou le tri d'une liste) ou est transformé en un format distinct (par exemple le codage historique en UTF-8). Une validation supplémentaire peut être réalisée lors du traitement.

Recommandations de l'UASG

- Étant donné qu'Unicode évolue, mettez à jour le logiciel lorsque vous le pouvez afin d'utiliser la dernière version de la norme et les éléments visuels et polices disponibles. N'oubliez pas qu'il se peut que les dispositifs utilisateurs, les bibliothèques de logiciels et les normes du web ne prennent pas en charge la dernière version, et puissent de ce fait afficher de manière erronée les caractères nouvellement attribués tels qu'une case générique (□), ou ne pas les afficher du tout.
- Lorsque les API qui prennent en charge les entrées et sorties UTF-8 sont disponibles, utilisez-les à la place des API ne les prenant pas en charge. Utilisez des bibliothèques standards bien déboguées, telles que la GNU libidn2 (https://www.gnu.org/software/libidn/#libidn2), afin de traiter et de valider les IDN ; n'utilisez pas la vôtre.
- Les scripts qui sont écrits de droit à gauche doivent faire l'objet de mesures particulières lorsqu'ils sont utilisés dans des noms de domaine et des adresses électroniques. Certaines de ces mesures particulières sont exposées dans l'IDNA[7] (pour les noms de domaine) et une annexe à la norme Unicode[8] (pour les adresses électroniques).
- Lors de la création de registres ou autres structures de données incluant des informations de script ou de langue, autorisez le plus possible ces informations, idéalement toutes celles que la norme Unicode prend en charge.[9] Gardez à l'esprit que certaines langues peuvent être écrites à l'aide de différents scripts et que certains scripts peuvent être utilisés afin d'écrire dans un grand nombre de langues.

[7]Voir le RFC 5893, « Scripts écrits de droite à gauche pour les noms de domaine internationalisés des applications (IDNA) » (https://tools.ietf.org/html/rfc5893).

[8]Voir l'UAX n° 9, « Algorithme bidirectionnel Unicode » (http://unicode.org/reports/tr9).

[9]Voir les « Scripts pris en charge par Unicode » (http://unicode.org/standard/supported.html).

AFFICHAGE

L'**affichage** intervient dès qu'une interface utilisateur fait apparaître une adresse électronique ou un nom de domaine. L'affichage des noms de domaine et adresses électroniques ne présente généralement pas de difficultés lorsque les scripts utilisés et les mécanismes d'affichage requis sont pris en charge par le système d'exploitation utilisé et que les chaînes sont stockées sous un codage défini par la norme Unicode. Sinon, des transformations propres à l'application peuvent être requises.

Recommandations de l'UASG

- Sachez que bien que les logiciels et dispositifs modernes puissent afficher presque tous les points de code Unicode, les anciens systèmes peuvent avoir une prise en charge limitée et exiger que les applications gèrent certaines de leurs anciennes polices. De même, lorsqu'Unicode ajoute de nouveaux points de code, les dispositifs et applications ne les afficheront pas jusqu'à la mise à jour de leurs bibliothèques de polices.
- Affichez les IDN sous leur forme de caractère d'origine sauf en cas d'obligation spécifique de les afficher en tant qu'étiquettes A.
- Les noms de domaine et adresses électroniques peuvent être affichés de gauche à droite (LTR), comme en anglais ou en russe, ou de droite à gauche (RTL), comme en arabe ou en hébreu. Dans la mesure où Unicode affecte des attributs directionnels aux points de code individuels (et pas aux séquences de points de code), certains textes mélangés LTR et RTL (« bidirectionnels ») peuvent être compris des utilisateurs, et d'autres non. Utilisez les critères de niveaux de restriction Unicode[10] afin de signaler des chaînes susceptibles de prêter à confusion.
- Les internautes lisent et parlent un grand nombre de langues. Dans certains cas, il peut être nécessaire de concevoir des applications distinctes pour les différentes langues et différents groupes de langues.

[10]Voir la norme technique d'Unicode n° 39, « Mécanismes de sécurité Unicode » (https://www.unicode.org/reports/tr39/#Restriction_Level_Detection), pour ses niveaux de restriction moyens et élevés afin de vérifier que les scripts d'une séquence Unicode respectent les bonnes pratiques.

Se préparer pour l'acceptation universelle

Révision de codes sources et tests unitaires

Les logiciels et systèmes qui ont été développés ou mis à niveau afin de prendre en charge l'acceptation universelle doivent être révisés et testés afin de s'assurer de leur exactitude et de détecter et résoudre tout bogue. Dans le cadre des initiatives de sensibilisation de l'UASG, le groupe cible les développeurs d'applications et les fournisseurs de services en ligne afin de les encourager à effectuer des révisions et des tests du code source d'acceptation universelle et à partager une liste de critères pouvant être utilisés afin de développer des tests types.

Tests

L'UASG développe également une liste de sites web, applications, adresses électroniques et noms de domaine pouvant être utilisés pour les tests. Dans certains cas, les tests peuvent être automatisés et mis en place sans intervention manuelle. Comme exemple concret, on peut citer l'enquête sur les gTLD récemment menée par APNIC Labs au nom de l'ICANN : https://tinyurl.com/new-gtld-ua. L'UASG cherche actuellement des méthodes de test automatisé pour l'acceptation universelle et partagera ses conclusions dès qu'elles seront disponibles.

Pour en savoir plus

Les documents suivants fournissent des informations supplémentaires sur l'acceptation universelle, Unicode et les noms de domaine internationalisés.

- UASG 007, « Introduction à l'acceptation universelle » (https://uasg.tech/documents).
- RFC 5894, « Noms de domaine internationalisés des applications (IDNA) : généralités, explications et fondements » (https://www.rfc-editor.org/info/rfc5894).
- « Typographie internationale sur le web », une synthèse graphique montrant les problèmes et enjeux liés à l'utilisation de plusieurs langues sur le web (https://w3c.github.io/typography/gap-analysis/language-matrix.html).

Remarque sur la terminologie

L'une des difficultés de l'acceptation universelle est que bon nombre de termes et de concepts connus des personnes habituées aux scripts simples ayant peu de caractères « alphabétiques » distincts, tels que le script latin, peuvent prêter à confusion lorsqu'ils sont appliqués à des systèmes d'écriture utilisant différents principes. L'intégration de toute une variété de systèmes d'écriture dans le domaine des noms de domaine internationalisés (IDN) a requis l'invention de nouveaux termes et le recours à des termes familiers (tels que « caractère ») dans des acceptions très spécifiques. Ce petit guide tâche d'éviter ces termes ou de les définir lorsqu'ils sont utilisés, mais l'examen d'autres supports, dont certains documents dont il est ici fait référence, nécessitera probablement une meilleure compréhension de la terminologie.