Universal Acceptance

Évaluation générale de l'acceptation des adresses électroniques par les sites Web en 2019

Entre
le Groupe directeur sur l'acceptation universelle (UASG),
l'Associação Brasileira das Empresas de Software (ABES)
et Governance Primer

9 août 2019

TABLE DES MATIÈRES

Introduction

L'objectif de l'acceptation universelle (UA) est de garantir que tous les noms de domaine et toutes les adresses de courrier électronique puissent être utilisés par toutes les applications, tous les dispositifs et tous les systèmes de l'Internet. Cela comprend tant les nouveaux noms de domaine génériques de premier niveau (gTLD) que les noms en scripts non latins. Malgré ce que l'on peut supposer, ces noms ne fonctionnent pas de la même manière que les noms historiques, et les problèmes de compatibilité restent plus fréquents qu'ils ne devraient l'être.

Notre objectif est que cette adresse de courrier électronique : **测试1@server.technology**	et cette adresse de courrier électronique : **دون@رسيل.السعودية**

aient le même taux d'acceptation que celle-ci :
user@test.org

Cette étude a été commanditée par le Groupe directeur sur l'acceptation universelle (UASG) pour faire suite à un test similaire effectué en 2017[1], dans le cadre d'une initiative plus large visant à sensibiliser la communauté aux obstacles et aux défis à relever pour parvenir à une compatibilité généralisée de tous les noms de domaine disponibles actuellement.

L'objectif était d'évaluer la conformité avec l'UA des 1000 principaux sites Web du monde (selon Alexa[2]) en faisant un échantillonnage des différentes pratiques et approches de développement utilisées pour le champ « courrier électronique » dans les formulaires Web, pour ensuite les tester dans la pratique. Ces formulaires pour courrier électronique s'avèrent utiles pour tester divers aspects de l'UA, étant donné qu'il est possible de vérifier différents points de défaillance, y compris les normes HTML et leur mise en œuvre, ainsi que l'utilisation de JavaScript et d'autres langages orientés Web.

Il est évident qu'un certain nombre de développeurs ne tiennent pas compte des nouveaux cas d'utilisation. En conséquence, les personnes qui souhaitent utiliser des adresses novatrices ou écrire dans leur propre langue doivent avoir une adresse de courrier électronique alternative à portée de main pour l'utiliser lorsque leur adresse préférée ne peut pas être traitée. Il est nécessaire de comprendre d'où viennent ces problèmes pour pouvoir y mettre fin.

Le reste du présent rapport fournira des informations sur la méthodologie de l'enquête, ses résultats, ses difficultés et des solutions potentielles.

[1] https://uasg.tech/wp-content/uploads/2017/09/UASG-Report-UASG017.pdf
[2] https://www.alexa.com/topsites

Résumé analytique

Après filtrage des noms de domaine faux, 527 des principaux 1000 sites Internet tirés du classement Alexa ont été identifiés comme susceptibles d'être testés. L'équipe a poursuivi sa recherche de sites Web pouvant faire l'objet de cette étude jusqu'à l'obtention de 1000 domaines « testables », ce qui correspondait à la position 1922 du classement Alexa. Des évaluations manuelles ont été effectuées pour vérifier l'acceptation des formats d'adresses suivants, par ordre de complexité ascendant, dans leurs formulaires :

ascii@ascii.newshort test@test.exp	ascii@ascii.newlong test@test.example	ascii@idn.ascii test@普遍接受-测试.org
unicode@ascii.ascii 测试1@test.org	unicode@idn.idn 测试5@普遍接受-测试.世界	arabic.arabic@arabic (RTL) دون@رسيل.السعودية

Les résultats ont ensuite été consignés dans une base de données MongoDB, suivant un schéma de métadonnées créé au cours du développement de la méthodologie afin de faciliter leur consultation. Les résultats présentés dans ce rapport sont le fruit de ce travail. L'analyse des écarts d'acceptation constatés entre les tests de 2019 et ceux de 2017 montre une tendance encourageante à la hausse de l'acceptation pour une certaine partie de la mission de l'UA, comme indiqué ci-dessous :

Cas de test	**2017**	**2019**
ascii@ascii.newshort	91%	97%
ascii@ascii.newlong	78%	84%
ascii@idn.ascii	45%	50%
unicode@ascii.ascii	14%	13%
unicode@idn.idn	8%	8%
arabic.arabic@arabic (RTL)	8%	7%

Certains écarts peuvent être attribués à l'utilisation de différents ensembles de données, étant donné que chaque test utilise la liste des principaux sites Web identifiés au moment de la collecte des données de chaque processus d'évaluation. Dans certains cas, il serait possible que l'acceptation de certains cas de test n'ait pas diminué, mais qu'elle soit restée figée. Toutefois, l'acceptation croissante de nouveaux domaines et noms de domaine internationalisés (IDN) au second niveau montre des progrès incontestables et revêt un grand intérêt.

Le principal problème identifié par notre équipe concerne le champ **<input type="email">**, qui en HTML5 correspond au champ de stockage des données à traiter, sur lequel repose un nombre considérable de sites Web. Le fait est qu'il n'est pas conforme à l'UA. L'amélioration de cette norme s'impose donc à nos yeux comme la principale priorité pour les parties prenantes concernées. Cela permettrait d'accroître l'acceptation de manière générale, surtout si les sites Web qui prennent en charge l'UA ont la possibilité d'utiliser un champ **<input type="eaiemail">** à la place, pour signaler ainsi leur capacité à accepter ce type d'adresses.

L’équipe de recherche

[Le superviseur] Paulo Milliet Roque, vétéran de l’industrie de la technologie, a de l’expérience dans le commerce international et a négocié avec plus de 100 entreprises dans plusieurs pays. Il est co-fondateur et vice-président d’ABES, l’Associação Brasileira das Empresas de Software (Association brésilienne des entreprises de logiciel). Fondée en 1986, ABES est l’entité la plus représentative du secteur avec environ deux mille entreprises associées et affiliées.

[Le coordinateur] Mark W. Datysgeld, ambassadeur de l’UASG, est titulaire d’une licence et d’un master en relations internationales, avec une spécialisation dans la gouvernance de l’Internet et les conséquences de la technologie sur l’élaboration de politiques publiques et privées. Sous l’enseigne Governance Primer, il conseille des sociétés et des personnes sur leur participation aux institutions et aux événements internationaux liés à la technologie.

[Le testeur principal] Sávyo Vinícius de Morais est membre de l’équipe NextGen de l’ICANN et ambassadeur du programme. Il mène des recherches dans le domaine de la sécurité et particulièrement sur la sécurité de l’Internet des objets dans l’environnement SOHO (petits bureaux et foyers), avec l’objectif d’atténuer l’impact d’une attaque massive sur les dispositifs IoT. Il a déjà participé à ce type de projets et travaille activement dans le développement Web et la gestion de systèmes.

[Le testeur] Edson Celio Ferreira Araujo est un jeune étudiant en informatique qui travaille comme développeur de systèmes à Grendene S/A. Dans son temps libre, il participe à des projets de code ouvert et collabore avec des initiatives liées à la gouvernance de l’Internet. Il est un ancien étudiant du programme Jeunesse@IGF.

[Le testeur] Jonas Mendes Fiorini est un jeune technicien en informatique, actuellement étudiant en ingénierie à l’Universidade Federal do Espírito Santo (UFES). Il participe à de nombreux projets portant sur l’inclusion numérique et sur des solutions de mise en réseau. Promoteur des logiciels libres depuis 2012, Fiorini est aussi un ancien étudiant du programme Jeunesse@IGF.

Nos précieux alliés sont : Don Hollander, Ajaay Data, Sarmad Hussain, Jennifer Chung, Nivaldo Cleto, Rodrigo de la Parra, Daniel Fink, Seda Akbulut, Vanda Scartezini, Rubens Kuhl, NIC.br.

Méthodologie

Procédures

Notre méthodologie est basée sur l'ouvrage « UASG017 : Évaluation de l'acceptation de diverses adresses de courrier électronique par des sites Web », une étude clé datant de septembre 2017 qui a répertorié pour la première fois la conformité avec l'UA à grande échelle. Cette étude de 2019 développe davantage les éléments essentiels de l'enquête précédente dans le but de faire évoluer ces méthodes.

Une importante décision méthodologique était de ne pas utiliser le même ensemble de sites Web que celui utilisé dans le test de 2017. Bien que la comparaison de sites web peut s'avérer intéressante pour évaluer une éventuelle évolution de leur taux de conformité, il ne s'agirait pourtant pas d'un état des lieux des principaux sites Web du monde entier mais plutôt d'une autre étude à part entière, qui serait tout aussi valable.

Il y a également des différences par rapport aux procédures sur lesquelles se fonde cette décision, y compris le fait que l'ensemble définitif de données du test précédent comportait environ 750 entrées contre 1000 de la version 2019. Il ne serait pas réaliste d'élargir rétroactivement l'ensemble de données de 2017, si bien qu'une nouvelle base de données est apparue comme l'option la plus raisonnable dans ce cas. Dans l'avenir, les deux options pourront être utilisées.

La première étape de l'étude, mise en place début 2019, a consisté à répertorier les URL sur lesquelles le test pouvait être effectué, qui correspondaient aux 1000 principaux sites Web identifiés grâce à l'outil d'analyse concurrentielle Alexa. Le choix des sites Web reposait sur certains critères : ils ne devaient pas contenir de logiciels malveillants et devaient comporter un champ de saisie d'adresses de courrier électronique quelque part dans leurs pages. Le contenu des sites n'a pas été pris en considération, raison pour laquelle tous les sites Web ont été traités comme étant également valables, quel que soit leur objet.

Seuls 527 des 1000 principaux sites Web ont satisfait aux critères du test, ce qui a poussé l'équipe à en chercher d'autres plus loin dans le classement. La recherche s'est arrêtée à la position 1922, qui correspond à la dernière entrée de l'ensemble de données. Si l'on devait définir de manière plus précise l'ensemble de données, on devrait dire qu'il s'agit de la liste des « 1000 principaux sites Web testables d'après Alexa ».

Pour se faire une idée de la portée géographique du test, l'équipe a également cherché à connaître la nationalité des sites Web à l'aide des registres WHOIS qui étaient, à l'époque, à jour. Cette étude a mis en évidence le fait qu'une grande partie des sites que l'on retrouve sur la liste correspond à des pays d'Occident, et notamment aux États-Unis. En Orient, la Chine, la Russie, le Japon, l'Inde, la Corée du Sud, le Taïwan et Hong Kong se distinguent par leur forte présence.

La répartition par codes de pays se trouve ci-dessous.

Veuillez noter que la somme n'atteint pas les 1000 en raison de l'impossibilité à identifier l'origine d'un petit nombre de sites Web.

AE	3	**CN**	47	**HK**	8	**LV**	1	**AS**	2	**VC**	1
AM	1	**CR**	1	**ID**	7	**MA**	2	**SC**	1	**VE**	1
AR	5	**CY**	8	**IE**	3	**MU**	1	**SE**	3	**VG**	4
AT	1	**CZ**	11	**IL**	6	**MX**	2	**SG**	2		
AU	13	**DE**	29	**IM**	1	**NG**	1	**SI**	1		
BA	1	**DK**	3	**IN**	16	**NL**	4	**SK**	1		
BD	1	**DO**	1	**IR**	7	**NO**	2	**TH**	2		
BE	1	**EC**	1	**IT**	14	**PA**	35	**TN**	1		
BG	1	**EG**	2	**JP**	24	**PE**	1	**TO**	3		
BR	13	**ES**	18	**KE**	1	**PH**	3	**TR**	9		
BS	10	**FR**	26	**KR**	11	**PL**	10	**TW**	12		
BY	1	**GB**	9	**KZ**	2	**PT**	4	**UA**	2		
CA	30	**GI**	1	**LA**	1	**RO**	2	**UK**	23		
CH	2	**GR**	3	**LU**	13	**RU**	33	**US**	381		

Six adresses de messagerie ont ensuite été créées pour pouvoir interagir avec les champs du formulaire des URL sélectionnés, chacune ayant des niveaux croissants de complexité par rapport aux domaines historiques standard en caractères ASCII. Les mêmes adresses ont été utilisées depuis le début de l'enquête jusqu'à sa conclusion. La liste des adresses de messagerie est la suivante :

Cas de test	Exemple
ascii@ascii.newshort	test@test.exp
ascii@ascii.newlong	test@test.example
ascii@idn.ascii	test@普遍接受-测试.org
unicode@ascii.ascii	测试1@test.org
unicode@idn.idn	测试5@普遍接受-测试.世界
arabic.arabic@arabic (RTL)	دون@رسيل.السعودية

Un écart important par rapport au test de 2017 a été l'abandon de l'adresse de messagerie « ascii@ascii.idn ». Au début, ce n'était pas l'intention de l'équipe, car cette messagerie faisait partie des cas d'utilisation prévus dans le RFC concerné. Cependant, après avoir tenté d'enregistrer sans succès une adresse de courrier électronique de test avec ce format auprès de cinq opérateurs de registre différents, et après avoir constaté que ce format n'était pas pris en charge, nous avons conclu avec les dirigeants de l'UASG que ce cas d'utilisation serait laissé de côté pour le test actuel car il n'était pas adopté de manière significative à l'heure actuelle.

Un élément hérité du test précédent était le cas de test « arabic.arabic@arabic » qui, représenté dans sa nomenclature correcte, serait un cas de test de droite à gauche (RTL). Nous avons corrigé dans une certaine mesure cette désignation erronée en ajoutant RTL à l'étiquette, mais nous nous sommes abstenus de changer complètement le nom en raison de l'absence actuelle de tests en hébreu, persan, ourdou, sindhi, yiddish et d'autres langues pertinentes. D'autres tests devront prendre cela en compte.

L'étape suivante de la procédure consistait à tester dans la pratique les formulaires destinés au courrier électronique en y insérant les adresses des cas de test, une à la fois, et en envoyant les formulaires aux sites Web chargés de les traiter. À ce stade, certains formulaires ont été ignorés en raison d'exigences de vérification par SMS ou de captchas impossibles à résoudre avec un clavier occidental normal. Pour ce deuxième cas de figure, même si le nombre de cas n'a pas été suffisamment élevé pour fausser les résultats, il est possible que cela puisse y avoir introduit un certain biais.

À chaque fois que ce processus était accompli avec succès, le site Web se voyait attribuer une catégorie parmi six possibilités entre Accepté ou Rejeté, suivant les critères suivants :

Marqué « Accepté » lorsque :

- √ L'envoi avait donné lieu à un message de confirmation.
- √ L'envoi avait été accepté et aucune erreur n'avait été signalée.
- √ Un message indiquant « courrier électronique déjà enregistré » était affiché[3].

Marqué comme « Rejeté » lorsque :

- ⊗ Le site Web affichait une erreur après la saisie de l'adresse.
- ⊗ L'envoi donnait lieu à un message d'erreur.
- ⊗ L'envoi n'était pas permis.

La table résultante avait le format suivant :

Site Web	E-mail 1	E-mail 2	E-mail 3	E-mail 4	E-mail 5	E-mail 6
test.org	Accepté	Accepté	Accepté	Rejeté	Rejeté	Rejeté
ページ.日本	Accepté	Accepté	Accepté	Accepté	Accepté	Rejeté

Le test des 100 premiers sites Web a pris à l'équipe 10 minutes par site en moyenne. À mesure que le test avançait, la moyenne de temps a diminuée à environ cinq minutes par site Web, ce que nous considérons comme une attente réaliste pour d'éventuelles futures tentatives de reproduire cette méthodologie.

À mesure que les tests étaient effectués, le code HTML de chaque page était enregistré localement afin d'en extraire les codes de validation du champ « courrier électronique » pour leur analyse. Ce processus a pris une période de temps variable, suivant le type de solution et la technologie utilisée. La validation effectuée côté serveur ou en HTML5 pouvait prendre une minute ou moins. Cependant, d'autres technologies tel que JavaScript pouvaient

[3] Lorsque le site Web avait une base de données partagée avec un site enregistré au préalable.

prendre jusqu'à 15 minutes par site Web, d'autant plus que le code était souvent « minifié »[4].

Au lieu d'enregistrer tout simplement ces résultats dans un tableur numérique, l'équipe a choisi de les organiser dans une base de données MongoDB, une plate-forme orientée document JSON qui n'a pas besoin d'être modélisée. Cette solution s'est avérée évolutive et légère. Elle a permis à tous les testeurs d'accéder à la version la plus récente des tests à tout moment. Les données étaient faciles à visualiser, même pour un utilisateur moins expérimenté. Dans l'ensemble, cette option nous a semblé meilleure qu'une base de données SQL ou un tableur en ligne partagé.

Après la création d'une base de données principale, MongoDB permet d'organiser les données dans un nombre illimité de « collections », l'équivalent des tables. Chaque collection contient tous les résultats liés à un test spécifique. Une fois qu'elles sont correctement désignées, les collections existent comme des ensembles de données et sont interopérables. En suivant la convention de nommage « ua_<portée>_<année> », elles peuvent toutes être stockées dans la même base de données, avec des noms faciles à lire pour l'homme.

Par exemple, la collection correspondant à la présente étude se nomme « ua_global_2019 », ce qui signifie que si l'année prochaine une nouvelle collection « ua_global_2020 » venait s'ajouter, elles pourraient coexister sans problèmes, et les chercheurs devraient être en mesure de faire des renvois entre leurs données. Il est important de noter que les noms des collections font la distinction entre majuscules et minuscules. Une future enquête menée au Mexique pourrait être nommée « ua_regional_Mexico_2019 », et ainsi de suite.

La dernière étape a consisté à consolider toutes les données et à convertir la base de données résultante en un fichier CSV qui serait traité et évalué afin d'élaborer le présent rapport. Tout fichier CSV respectant le schéma de métadonnées détaillé ci-dessous peut être facilement importé dans la base de données à l'aide de la commande suivante, qui suppose un environnement Ubuntu :

```
mongoimport --db ua_database --collection <scope>_<year> --host=<hostname> --username=<username> --password=<password> --drop --type CSV --file <file_address>/ua_global_2019.csv –headerline
```

Schéma de métadonnées

Après avoir choisi MongoDB pour le stockage des données, l'équipe a établi un schéma de métadonnées que nous espérons sera employé pour des tests futurs, afin que les résultats des différents échantillons puissent tous partager la même norme et être comparés quelle que soit l'approche adoptée par les chercheurs.

Nous présenterons et expliquerons maintenant cette norme.

[4] Définition : Supprimer les données inutiles ou redondantes sans affecter la façon dont la ressource est traitée par le navigateur. Cela peut concerner la suppression du formatage ou des commentaires relatifs au code, la suppression du code non utilisé, l'utilisation des noms de fonction et de variable plus courts, etc. [Wikipédia]

```
{
   "_id": {"$oid": "00001"}
   "domain": "test.org",
   "url":
"https://www.test.org/signup",
   "rank": "1000",
   "testable": "Yes",
   "code": "<input
type="email">",
   "comments": "Field triggers
a captcha",
   "mailboxes":
   {
        "mail1": "Accepted",
        "mail2": "Accepted",
        "mail3": "Accepted",
        "mail4": "Rejected",
        "mail5": "Rejected",
        "mail6": "Rejected"
   }
}
```

Voici les fonctionnalités de chaque chaîne :

- **_id** : Identificateur unique auto-généré pour l'ensemble de la base de données.
- **domain** : Domaine, sans le préfixe « www ».
- **url** : URL comportant le parcours complet pour arriver à la page contenant le formulaire.
- **rank** : Classement du site Web dans cette collection spécifique.
- **testable** : Indique si l'équipe a finalement réussi à tester le site Web.
- **code** : Contient la chaîne qui valide le formulaire, si elle est trouvée.
- **comments** : Tout commentaire pertinent.
- **mailboxes** : Liste des classements de chaque cas de test.

Pour des tests futurs, nous prévoyons l'inclusion d'un champ **« eai »**, qui permettrait de déterminer si un serveur de messagerie donné prend en charge l'internationalisation des adresses de courrier électronique, ce qui signifie fondamentalement qu'il serait en mesure d'échanger des messages en UTF-8 et par conséquent en Unicode. Cela nous permettrait de mieux connaître le niveau général de prise en charge de cette technologie, grâce à l'identification du nombre de points de défaillance rencontrés avant de pouvoir échanger avec succès des messages à partir d'adresses plus complexes.

Pour ce rapport 2019 en particulier, une chaîne « **pays** » avait été incluse afin de pouvoir déterminer un peu plus clairement la portée géographique de la liste. Cependant, l'inclusion de cette chaîne n'est pas proposée comme un élément constitutif de la norme en raison des modifications apportées à la base de données WHOIS qui empêchent des recherches massives comme celle effectuée par l'équipe et qui rendraient donc impossible ce type de

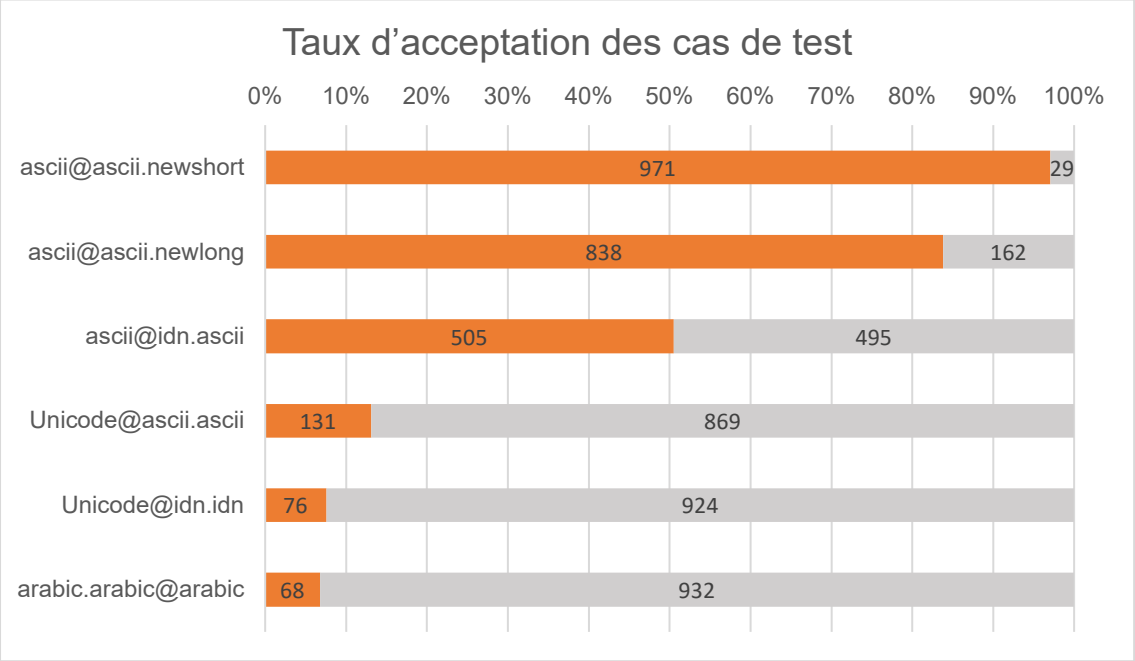
pratique. Si le RDAP acceptait finalement de telles requêtes, l'inclusion de la chaîne pourrait à nouveau être considérée.

Résultats

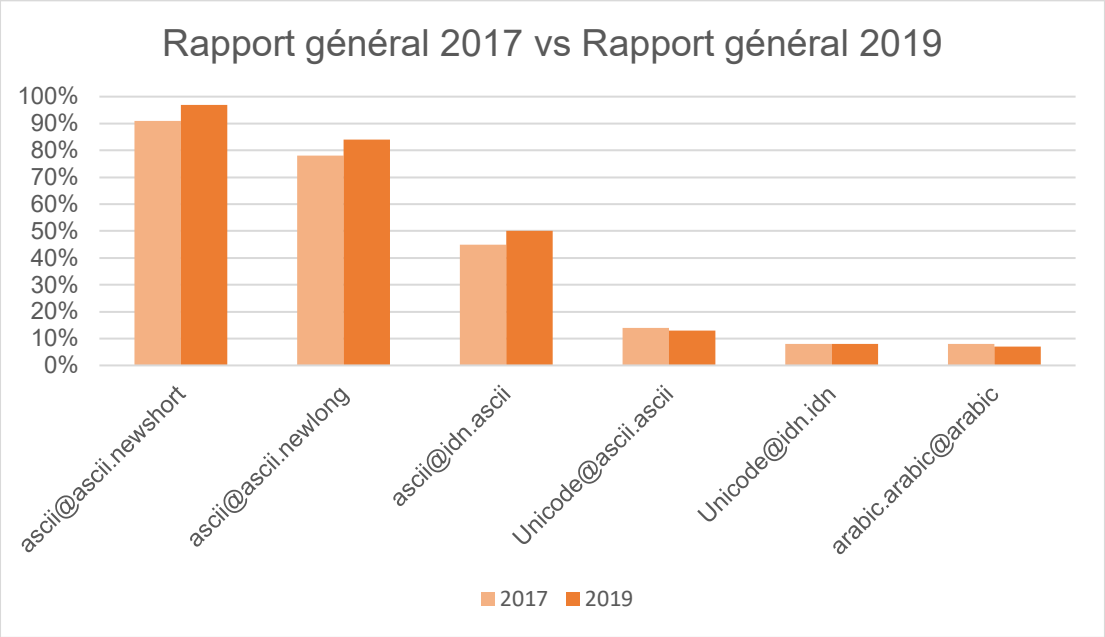
Résultats des tests en chiffres

Cas de test	Accepté	Rejeté
ascii@ascii.newshort	971	29
ascii@ascii.newlong	838	162
ascii@idn.ascii	505	495
Unicode@ascii.ascii	131	869
Unicode@idn.idn	76	924
arabic.arabic@arabic (RTL)	68	932

Graphique des résultats des tests



Comparaison des résultats des tests de 2017 et de 2019



Évaluation et corrélations

Dans ce chapitre nous passerons en revue certains résultats que l'équipe a rencontrés au cours de l'évaluation des codes, qui permettent d'identifier les mauvaises pratiques autour de l'utilisation du formulaire pour courrier électronique sur le Web. Nous avons aussi profité de l'occasion pour mener une expérience à petite échelle où l'on met à profit la collecte de données régionales.

Il est important de signaler que, tout comme cela a été noté dans l'enquête de 2017, il n'existe pas d'approche unifiée pour le codage des fonctions de validation ailleurs que dans le cas du HTML5 et de quelques scripts normalisés, tels que « jquery-validation.js ». Un nombre important de sites Web s'appuient sur l'utilisation d'expressions régulières (RegEx) en JavaScript, avec des résultats plus ou moins satisfaisants. Or, en termes généraux, leur usage est rarement une bonne solution.

Échecs de validation des échantillons

Les sites Web utilisant les meilleurs codes mais affichant un faible taux de conformité à l'UA rejetaient les adresses de courrier électronique dès qu'elles étaient saisies dans le formulaire. Les sites Web avec des codes moins efficaces acceptaient d'emblée les adresses mais finissaient par afficher un message d'erreur (« quelque chose a mal tourné »). Dans un petit nombre de cas, le site Web a identifié les cas de test les plus complexes comme étant malveillants et les a classés comme des attaques ou des tentatives de piratage.

Voici quelques exemples des messages souvent rencontrés, à quelques variantes près, lorsqu'un cas de test est rejeté :

Veuillez saisir une adresse de courrier électronique valide.
Format de courrier électronique invalide.
Désolés, cette adresse de courrier électronique ne semble pas correcte. Veuillez vérifier que ce soit bien une adresse de courrier électronique.
Le texte avant le caractère « @ » ne devrait pas contenir des symboles.
Quelque chose a mal tourné.

Un type de message en particulier se détache : celui indiquant qu'il y a un problème si le caractère @ est précédé par des symboles. Ce message est apparu régulièrement dans toutes nos observations et il semblerait être la principale source d'erreur à l'origine des problèmes de validation.

Rejet de toutes les adresses de courrier électronique

L'équipe a trouvé quelques exemples particulièrement mauvais de pratiques de codage parmi les 1000 principaux sites Web du monde. Nous reprenons ci-dessous un exemple en JavaScript qui n'a admis aucun cas de test et qui correspond, en plus, à un site Web vietnamien. Il ne s'agit pas d'ailleurs d'un cas isolé ; d'autres sites Web dans la liste répètent la même erreur de diverses manières.

RegEx :

```
/^[-a-z0-9~!$%^&*_=+}{\'?]+(\.[-a-z0-9~!$%^&*_=+}{\'?]+)*@([a-z0-
9_][-a-z0-9_]*(\.[-a-z09_]+)*\.(aero|arpa|biz|com|coop|edu|gov|
```

```
info|int|mil|museum|name|net|org|pro|travel|mobi|[a-z][a-z])|([0-
9]{1,3}\.[0-9]{1,3}\.[0-9]{1,3}\.[0-9]{1,3}))(:[0-9]{1,5})?$/i
```

Cette partie du code effectue une vérification rigoureuse de certains paramètres très spécifiques qui définissent la structure d'une adresse de courrier électronique historique, tout en prenant le soin d'inclure dans une liste blanche (whitelist) certains TLD sponsorisés et, de manière redondante, des TLD historiques tels que « .edu » et « .org ». Cela n'a pas beaucoup de sens, mais des variantes de cette pratique ont été retrouvées dans d'autres instances.

Tests de corrélation régionale

Bien que ce ne soit pas un des objectifs principaux de l'étude, l'équipe a essayé d'établir une corrélation entre le taux de conformité avec l'UA des principaux sites Web et leur appartenance à des pays qui utilisent des scripts autres que le latin, en espérant que les niveaux de conformité seraient plus élevés. Après avoir étudié les scripts cyrillique et han, nous avons conclu que **cette corrélation semble ne pas exister** à présent. De futurs tests menés au niveau local pourront peut-être mieux évaluer si sur les 100 principaux sites au niveau régional les résultats sont meilleurs, ce qui représente un enjeu important pour l'avenir.

Voici les résultats :

- **Script Han** : l'échantillon était composé de 47 sites Web de Chine, 8 de Hong Kong, 2 de Singapour et 12 de Taïwan. Au total, seuls 4 sites Web de Chine ont pu traiter le cas de test relativement simple « unicode@ascii.ascii ».
- **Script cyrillique** : l'échantillon était composé de 37 sites Web de Russie et 2 d'Ukraine. Au total, seuls 5 sites Web russes ont pu prendre en charge le cas de test relativement simple « unicode@ascii.ascii ».

HTML5 : l'obstacle à l'acceptation universelle

Les résultats de l'étude « ua_regional_Brazil_2018 »[5] ont montré que 30 pour cent des 50 principaux sites Web du pays utilisaient le champ **<input type="email">** comme solution pour la validation des adresses de courrier électronique côté navigateur. L'équipe s'attendait à trouver une tendance similaire dans l'enquête mondiale, et son hypothèse s'est confirmée. Le taux d'utilisation mondiale de cette chaîne varie entre 20 et 30 pour cent selon les tendances que nous avons observées.

Le problème est que mi-2019 la norme HTML5 n'est toujours pas adaptée à l'acceptation universelle. Cela nous préoccupe, d'autant plus que le déploiement de la norme ne cesse de progresser et que les développeurs s'appuient de plus en plus sur ses fonctionnalités pour faire fonctionner leurs codes dans les nombreux cas de figure que l'on retrouve dans l'état actuel de la demande en matière de sites Web.

Voici le modèle actuel d'acceptation pour **<input type="email">** :

Cas de test	Résultat
ascii@ascii.newshort	Accepté
ascii@ascii.newlong	Accepté
ascii@idn.ascii	Accepté
unicode@ascii.ascii	Rejeté
unicode@idn.idn	Rejeté
arabic.arabic@arabic (RTL)	Rejeté

Ce résultat ne fait que renforcer notre affirmation précédente, à savoir que la principale source d'erreur viendrait de l'hypothèse selon laquelle il ne devrait pas y avoir de symboles devant le caractère « @ ».

Si 20 à 30 pour cent des sites Web utilisent cette chaîne, cela veut donc dire que le même nombre de sites échoue à cette étape de la validation en raison d'un manque de normes HTML5 appropriées. Autrement dit, côté navigateur, il s'agit de la principale mesure qui peut être prise actuellement pour faire augmenter l'UA.

Après avoir examiné l'état de la question, nous avons constaté que la version préliminaire 5.3 de la spécification, dans son chapitre consacré à la validation de courriers électroniques[6], prévoit la conformité avec les RFC 6531 et RFC 5890, ce qui rendrait la norme conforme à l'UA. Toutefois, cette révision reste au stade de « première version de travail » depuis 2017, sans qu'il y ait une date prévue pour qu'elle devienne une « recommandation candidate » ou une « recommandation » finale.

---

5 http://central.abessoftware.com.br/Content/UploadedFiles/Arquivos/Estudos,%20Pesquisas%20e%20Pareceres/Universal-Acceptance-in-Brazil-2018.pdf

6 https://www.w3.org/TR/html53/sec-forms.html#email-state-typeemail

Nous trouvons que cela est préoccupant, compte tenu de l'enjeu que représente l'acceptation universelle pour la communauté du fait de son lien avec les noms internationalisés. La communauté ne peut pas rester les bras croisés : des mesures générales doivent être adoptées dans notre relation avec le W3C, WHATWG et les fournisseurs de navigateurs afin d'accélérer le processus de mise en œuvre de cet aspect de la spécification.

Cette démarche est d'autant plus opportune que le W3C et WHATWG sont récemment convenus d'adopter une approche commune pour l'élaboration de normes[7]. Il est important de contacter les parties concernées et de leur demander qu'une option « eaimail » soit ajoutée pour les sites qui sont adaptés à l'UA, ainsi que d'exercer des pressions sur les principaux fournisseurs de navigateurs afin qu'ils s'intéressent à faire en sorte que cette option soit viable.

L'équipe de recherche recommande que la plus haute priorité soit accordée à la résolution de ce problème.

[7] https://www.w3.org/blog/2019/05/w3c-and-whatwg-to-work-together-to-advance-the-open-web-platform/

Ces rapports peuvent-ils être automatisés ?

Au cours de l'élaboration de la présente étude, lors des réunions de l'ICANN et au cours des sessions de planification, des questions ont été soulevées sur la faisabilité d'un logiciel automatisé pour l'élaboration de rapports comme celui-ci. L'équipe a essayé de trouver une réponse pratique à la question, mais a conclu qu'il n'est pas possible d'automatiser ces tests pour le moment.

Voici nos principales préoccupations :

- Il existe une trop grande diversité de codages des sites web, ainsi que de scripts utilisés et de manières d'organiser les formulaires. Dans l'ensemble, il serait pratiquement impossible de prendre en compte toutes les variantes.
- La minification est largement utilisée pour réduire les temps de chargement des sites Web et des scripts, ce qui oblige parfois à faire appel à un expert en codage pour lire les résultats.
- Un nombre important de sites Web ont recours à des types de Captcha ou à certaines nuances dans leurs formulaires qui rendent l'automatisation inutile.

Une partie de ces arguments peut aussi s'appliquer à la question de savoir s'il est possible d'externaliser l'étude auprès de non-experts, comme la proposition d'utiliser le « Mechanical Turk » d'Amazon. Bien que cela puisse fonctionner en termes d'exécution du test, il serait toutefois nécessaire de coordonner la préparation de la liste des domaines, l'extraction du code et l'élaboration d'un rapport approfondi qui représente plus qu'un ensemble de statistiques. Cependant, cette option est bien plus faisable que l'automatisation.

Conclusion

Sur les 100 principaux sites Web du monde que nous avons pris comme référence, seuls cinq ont pris en charge tous les cas faisant l'objet du test : « quora.com », « espn.com », « spotify.com », « txxx.com » [NSFW] et « godaddy.com ». Si nous ne comptions pas le cas le plus complexe « unicode@idn.idn » et les cas de test en arabe (RTL), nous nous retrouverions toujours avec seulement 9 sites Web qui prennent en charge les quatre premiers cas de test. Puisque nous parlons des pages les plus consultées au monde, cela est tout à fait préoccupant.

Même en étant optimistes par rapport à la possibilité que les « newshorts » soient pleinement conformes dans un avenir proche et en considérant réaliste la possibilité que les « newlongs » suivent la même tendance, il est clair que les domaines internationalisés ont besoin d'un coup de pouce et qu'il est nécessaire de prendre des mesures proactives pour qu'ils prospèrent. L'élaboration de politiques plus robustes doit s'accompagner d'une sensibilisation proactive qui permette de faire connaître l'existence de ce type de domaines et adresses de courrier électronique.

En conclusion, nous considérons que la prochaine étape importante serait l'élaboration d'un tableau de bord permettant d'évaluer tous les rapports sur l'UA qui sont générés suivant le schéma de métadonnées proposé dans ce document. Cela contribuerait à créer un référentiel durable qui pourrait être utilisé par les parties intéressées pour effectuer des mesures réalistes et qui servirait d'outil de référence pour l'ensemble de la communauté.