Universal Acceptance

# Global Evaluation of Websites for Acceptance of E-mail Addresses in 2019

Between
The Universal Acceptance Steering Group (UASG),
Associação Brasileira das Empresas de Software (ABES),
& Governance Primer

9 August 2019

## TABLE OF CONTENTS

## Introduction

The goal of Universal Acceptance (UA) is to ensure that every domain name and e-mail address can be used by all Internet-enabled applications, devices, and systems. This contemplates both new generic top-level domains(gTLDs) and non-Latin-based ones. While it may be assumed by some that these work in the same manner as legacy ones, that is not the case, and problems with compatibility are still more common than they should be.

| Our goal is for this e-mail: | And this e-mail: |
| --- | --- |
| **测试1@server.technology** | **السعودية.رسيل@دون** |

To have the same rate of acceptance as this one:
**user@test.org**

This survey was commissioned by the Universal Acceptance Steering Group (UASG) as a follow-up to a similar 2017 test[1], as part of a broader initiative to further the community's understanding of the bottlenecks and key issues surrounding widespread compatibility of all domain names currently available.

The goal was to evaluate the UA compliance of the top 1,000 websites in the world (according to Alexa[2]) by sampling the practices and different development approaches to the e-mail field in forms on the Web and testing them in practice. These e-mail forms are well-suited to test diverse aspects of UA, as different fail points can be checked including HTML standards and implementation, as well as usage of JavaScript and other Web-oriented languages.

It is apparent that a number of developers do not take into consideration newer use-cases. As a result, people who want to use innovative addresses or be able to type in their own languages, need to have a fallback e-mail on hand for when their preferred address inevitably fails to be processed. There is a need to understand where these problems exist so they can ultimately be fixed.

The rest of this report will provide information on the methodology of the survey and outline results, issues, and potential solutions.

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

## Executive Summary

After filtering for bogus domains, 527 of the top 1,000 websites collected from Alexa's list proved to be testable. The team proceeded looking for testable websites until position 1.922 of the list, at which point a thousand testable domains were obtained. Manual assessments were then performed to check for the acceptance of the following and increasingly complex address formats in their forms:

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

Results were then committed to a MongoDB database for easier reference in a metadata schema created during development of the methodology, and the results presented on the report derive from that work. Taking into consideration the difference in acceptance between the 2019 and the 2017 tests, we can observe an encouraging pattern of increased acceptance for a certain portion of the UA mission, as can be observed below:

| Test case | 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% |

Some of the variance can be accounted for due to the use of different datasets, with each test using the list of the top websites as it existed during the data collection phase of each evaluation process. In this case, it might be that acceptance of some test cases might not be decreasing, but rather has remained at a standstill. However, the increase in acceptance of new domains and Internationalized Domain Names (IDNs) at the second level shows unequivocal progress and is a matter of great interest.

The most significant issue found by us was **<input type="email">**, which is HTML5's stock field for processing of such data, on which a significant number of websites rely on. The fact is: it is not compliant with UA. We believe the number one priority to concerned stakeholders should be the improvement of this standard. This would increase acceptance across the board, particularly if UA Ready websites could opt to use an **<input type="eaiemail">** instead, signaling their capacity to accept these addresses.

## The Research Team

**[Overseer] Paulo Milliet Roque** is a technology industry veteran with experience in international trade, who has negotiated with over 100 companies in several different countries. He is the co-founder and vice president of ABES, the Brazilian Software Association. Founded in 1986, ABES is the most representative entity in the sector with approximately two thousand associated and affiliated companies.

**[Coordinator] Mark W. Datysgeld** is a UASG Ambassador that holds a BA and Master of International Relations, focused on Internet Governance and the impacts of technology on public and private policymaking. Under the Governance Primer brand, he consults for businesses and individuals in their participation in international institutions and events that relate to technology.

**[Lead tester] Sávyo Vinícius de Morais** is an ICANN NextGen and Ambassador, researching security with a focus on IoT in SOHO (Small Offices and Home) environments with the objective of mitigating the impact of massive attacks on IoT devices. He has work experience and actively engages with Web development and systems administration.

**[Tester] Edson Celio Ferreira Araujo** is a young student of Computer Engineering that acts as a systems developer at Grendene S/A. In his free time, he contributes to open source projects and works in Internet Governance endeavors. He is an alumnus of the Youth@IGF program.

**[Tester] Jonas Mendes Fiorini** is a young technician in computing, currently studying Computer Engineering at Universidade Federal do Espírito Santo (UFES). He participates in projects related to digital inclusion and networking solutions. and has been a free software enthusiast since 2012. Fiorini is an alumnus of the Youth@IGF program.

**Our valued supporters are**: Don Hollander, Ajaay Data, Sarmad Hussain, Jennifer Chung, Nivaldo Cleto, Rodrigo de la Parra, Daniel Fink, Seda Akbulut, Vanda Scartezini, Rubens Kuhl, NIC.br.

## Methodology

### Procedures

The basis of our methodology comes from “UASG017: Evaluation of Websites for Acceptance of a Variety of Email Addresses,” a key study from September 2017 that mapped compliance with UA on a large scale for the first time. This 2019 study further elaborates on the essential components of the previous investigation with the intention of advancing those methods.

An important methodological decision made was that the pool of websites from the 2017 test was not used. While there is significant value in comparing whether or not websites progressed in their rate of compliance, the fact remains that it would not be an evaluation of the state of the current top websites in the world, rather being a different study altogether; one that is just as valid.

There were also some divergences in the procedures that inform this decision, including the fact that the final dataset of the previous test contained around 750 entries, while the 2019 version reached 1,000. It would be unrealistic to retroactively fit more entries into the 2017 dataset, so a fresh database was the option that made the most sense in this case. Moving forward, either option can now be pursued.

The first step taken in this study was to list the testable URLs of the top 1,000 websites in the world according to the competitive analysis tool Alexa, which was performed in early 2019. The selected websites had to meet the criteria of not being malware and of having an e-mail input field available somewhere in their pages. Questions of content were not accounted for so all websites were treated as equally valid regardless of their subject matter.

From the initial list of top 1,000 websites, only 527 of them met the testing criteria, prompting the team to proceed further down the list to find more candidates, eventually arriving at 1,922, which became the final entry in the dataset. Were we to be more precise about naming the dataset, it would be the list of the “top 1,000 testable websites according to Alexa.”

The team also looked up the nationality of the websites using a then still up-to-date WHOIS record, to get an impression of what the geographical reach of the test was. This pointed to the fact that much of the list is still concentrated around the Western Hemisphere, particularly dominated by the United States. In the Eastern Hemisphere, China, Russia, Japan, India, South Korea, Taiwan, and Hong Kong stand out as having the strongest presences.

The complete breakdown by country code follows.

Please note, the sum does not reach 1,000 due to the inability to detect the origin of a small number of websites.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **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 | | |

A set of six mailboxes were then created in order to interface with the form fields of the selected URLs, with each having increasing levels of complexity in relation to the standard ASCII-based legacy domains. The same addresses were maintained from the start to the conclusion of the survey. The list of mailboxes is as follows:

| Test Case | Example |
|---|---|
| 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 |

There was a significant departure from the 2017 test in the discarding of the “ascii@ascii.idn” mailbox. At first, this was not the team’s intent, as according to the relevant RFCs it is supposed to be a supported use case. However, after attempting to register a test case e-mail using that format on five different registries that all resulted in failures because the format was not supported, we concluded with UASG leadership that this would be left out for the current test as it is not being adopted in a significant manner at the moment.

One element inherited from the previous test was the “arabic.arabic@arabic” test case, which, if represented using its proper nomenclature, would be the Right-to-left (RTL) test case. We have ammended this misnomer to some degree by adding RTL to the label, but a

lack of testing of Hebrew, Persian, Urdu, Sindhi, Yiddish, and other relevant languages stopped us from transitioning the name completely. Further tests need to take this into consideration.

The next step in the procedure was the practical test of the e-mail forms by inserting the address of the test cases one at a time and submitting the forms to be processed by the websites. Some of the forms were skipped at this stage due to requirements of SMS verification or unsolvable captchas on a Western-standard keyboard. In this second case, while the number was not high enough to throw the results off, there is the potential that it might have tilted the survey to some degree.

Each time this process was successfully accomplished, the website was assigned six ratings of Accepted or Rejected, according to the following criteria:

Marked as Accepted when:

- √ Submission resulted in a success message.
- √ Submission was accepted and no error was reported.
- √ An “already registered” e-mail message was displayed[3].

Marked as Rejected when:

- ⊗ The website displayed an error once the address was inputted.
- ⊗ Submission returned an error message.
- ⊗ Submission was not allowed.

The resulting table had the following format:

| Website | E-mail 1 | E-mail 2 | E-mail 3 | E-mail 4 | E-mail 5 | E-mail 6 |
|---|---|---|---|---|---|---|
| test.org | Accepted | Accepted | Accepted | Rejected | Rejected | Rejected |
| ページ.日本 | Accepted | Accepted | Accepted | Accepted | Accepted | Rejected |

While testing the first 100 websites, it took the team an average of 10 minutes to complete each one. As the test progressed, the average time dropped to approximately five minutes per website, which we consider to be a realistic expectation for future attempts at repeating this methodology by anybody.

As the tests were performed, the HTML code of each page was saved locally so that the validation codes from the e-mail field could be extracted for analysis. This process took a variable amount of time, depending on the type of solution and the technology being employed. If the validation was performed server-side or in HTML5, that could be identified in one minute or less. However, other technologies such as JavaScript could take up to 15 minutes per website, more so because the code was often minified[4].

[3] When it happened that the website had a shared database with a previously registered one.
[4] Def.: Removing unnecessary or redundant data without affecting how the resource is processed by the browser, such as code comments and formatting, removing unused code, using shorter variable and function names, and so on. [Wikipedia]

Instead of simply committing these results to a digital spreadsheet, the team opted to organize it on a MongoDB database, a platform oriented towards JSON that doesn't need to be modelled. This solution proved to be scalable, light, and every tester had access to the most up to date version of the tests at any given time. It was easy to visualize the data and even a less experienced user could have operated it. Overall, we considered it a superior option to an SQL database or a shared online spreadsheet.

Once a master database is created in MongoDB, it allows for unlimited "collections," which are analogue to tables. Collections contain all the results pertinent to a particular test, and once properly named, they exist as self-contained datasets that are interoperable. By following the naming convention "ua_<scope>_<year>", all of them could be collected in the same database maintained in human-readable names.

For example, the collection for this study is named "ua_global_2019", meaning that if the one for the following year is named "ua_global_2020", they could coexist without problems, and researchers would be able to cross-reference their data. It is important to note that collection names are case sensitive. A future survey carried out in Mexico could be named "ua_regional_Mexico_2019", and so on.

The final step was then aggregating all the data, converting the resulting database into a CSV file, and taking it for processing and evaluation so that this report could be produced. Any CSV that follows the metadata schema detailed below can be imported to the database with ease with the following command, which supposes an Ubuntu environment:

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

### Metadata Schema

With MongoDB chosen for storage of the data, the team came up with a metadata schema that we hope is employed by future tests, so that results from different samplings can all share the same standard and be compared at will under whatever lens researchers choose.

We will now present this standard and explain it.

```
{
   "_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"
   }
}
```

These are the functionalities of each string:

- **_id**: Auto-generated unique identifier within the entire database.
- **domain**: Domain, not prefixed by “www”.
- **url**: URL containing the full path to the page containing the form.
- **rank**: Ranking of the website within that particular collection.
- **testable**: Indicates if the team ultimately managed to test the website.
- **code**: Contains the string that validates the form, if found.
- **comments**: Any relevant comment.
- **mailboxes**: List of ratings for each test case.

For future tests, we contemplate the inclusion of an “**eai**” field, that would indicate whether a given domain’s mail server supports Email Address Internationalization, which fundamentally means that they would be able to exchange mail in UTF-8 and consequently in Unicode. This would provide us with more insight into what is the general level of support for this technology, considering how many failure points there are on the way to the successful exchange of messages from more complex addresses.

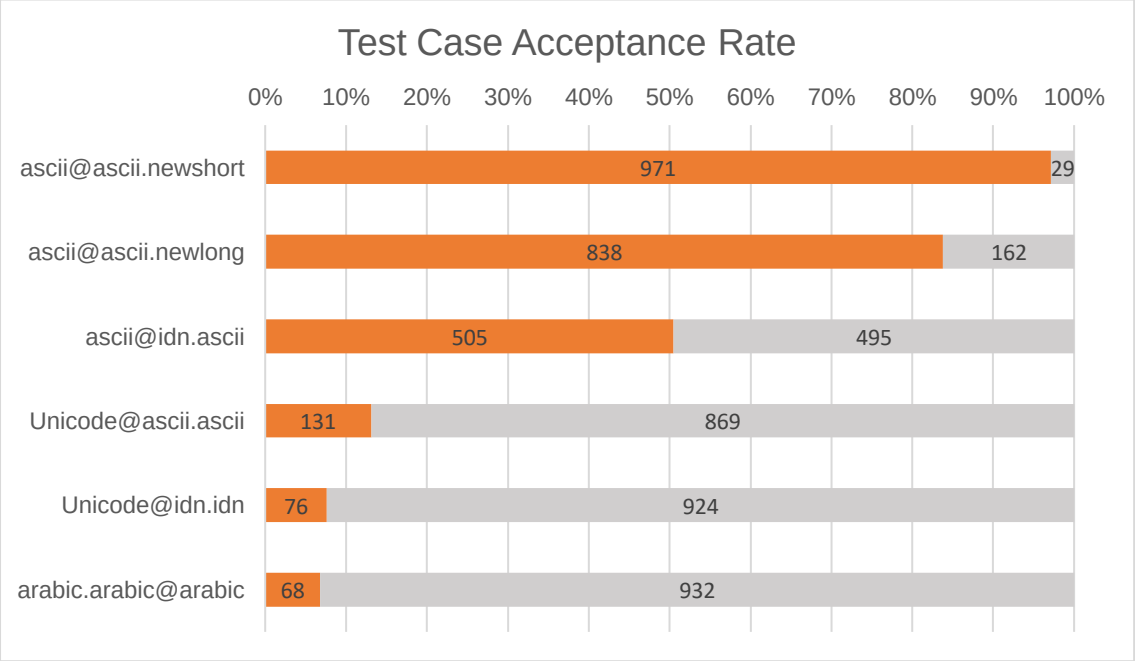
For this particular 2019 report, a “**country**” string had been included so that a better sense of the geography of the list could be assessed. However, this is not suggested as being part of the standard due to the changes made to the WHOIS database that prevent massive lookups such as the one performed by the team, which would make it impractical moving forward. If RDAP eventually allows for such queries to be made, the string might be made viable again.

## Results

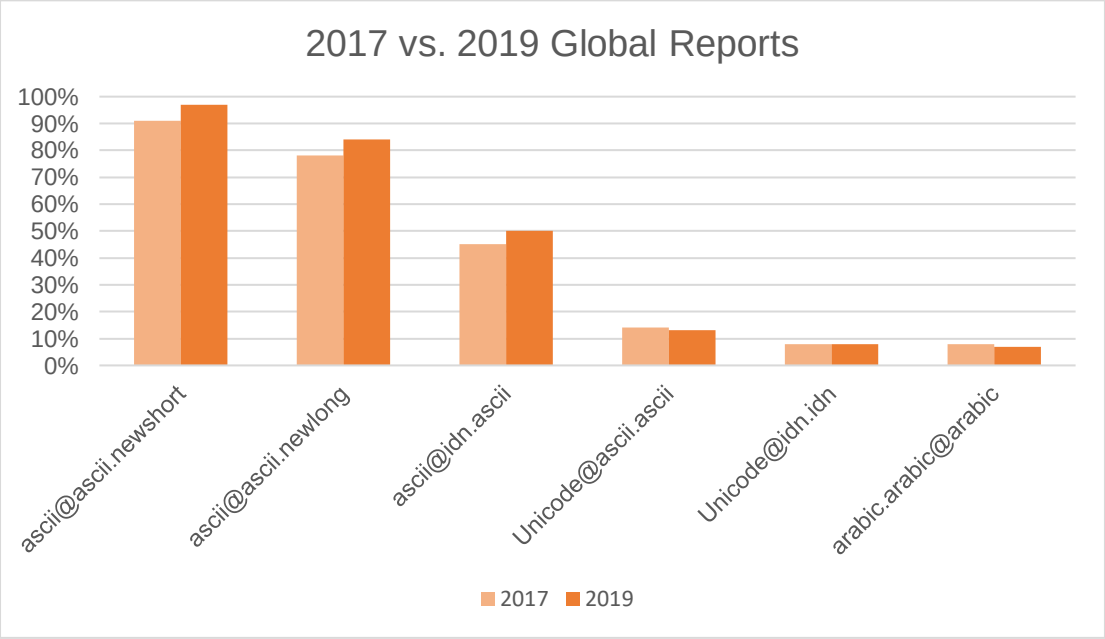
### Test Totals in Numbers

| Test Case | Accepted | Rejected |
|---|---|---|
| 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 |

### Test Totals in a Graph



**2017 vs. 2019 Test Totals Comparison**



## Evaluation and Correlations

In this section we will go through some of the results the team encountered during the evaluation of the codes, bringing to light what are the bad practices that surround e-mail form usage on the Web. We also took the opportunity to carry out a small-scale experiment taking advantage of the collection of regional data.

One important note to make is that much like what was noted in the 2017 survey, there is no unified approach to the coding of validation functions other than in the case of HTML5 and a few standardized scripts, such as “jquery-validation.js”. A significant number of websites rely on Regular Expressions (RegEx) in JavaScript, achieving varied degrees of success in their use. But as a rule of thumb, those are rarely a good solution.

### Sample Validation Failures

Websites with better code overall but poor UA compliance denied the e-mail outright as it was inputted into the form. Websites with poorer code accepted the address at a first glance but proceeded into a “something went wrong” message. In a few cases, the website identified the more complicated test cases as malicious, classifying them as attacks or hacking attempts.

These are some of the commonly encountered messages, and small variants thereof, when a test case was rejected:

| |
|---|
| Please enter a valid e-mail address. |
| Invalid e-mail format. |
| Sorry, that e-mail doesn’t look right. Please check it's a proper e-mail. |
| Text before the character @ should not contain symbols. |
| Something went wrong. |

One type of message in particular stands out – that there is a problem if the @ character is preceded by symbols. This was quite consistent across our observations and that seems to be the key misconception that is causing problems in terms of validation.

### Rejection of All E-mail Addresses

The team came across some particularly bad examples of coding practices within the top 1,000 websites in the world. Below we separated one example in JavaScript that failed every test case, which comes from a Vietnamese website no less. It is not an isolated sample either, and other websites in the list make the same mistake in a broad variety of ways.

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
```

What this piece of code does is a hard check for very specific parameters that define the structure of a legacy e-mail address, taking extra care to whitelist some of the Sponsored TLDs, as well as redundantly whitelisting legacy TLDs such as “.edu” and “.org”. It does not make much sense, but variations of this were also found in other instances.

## Regional Correlation Tests

While not a main goal of the study, the team took interest in attempting to correlate UA compliance with top websites from countries that employ writing scripts different from Latin, with the expectation that compliance levels might be higher. After studying both the Cyrillic and Han scripts, we concluded that **this correlation does not seem to exist** at this level. Perhaps future local tests can better assess if at the top 100 regional level, the websites fare better, which is an important question moving forward.

The results can be observed below:

- **Han case**: the pool was comprised of 47 websites from China, 8 from Hong Kong, 2 from Singapore, and 12 from Taiwan; on total, only 4 websites from China were able to handle the mild “unicode@ascii.ascii” test case.
- **Cyrillic case**: the pool was comprised of 37 websites from Russia and 2 from Ukraine; on total, only 5 Russian websites were able to handle the mild “unicode@ascii.ascii” test case.

## HTML5: The Roadblock to Universal Acceptance

Back in “ua_regional_Brazil_2018”[5], the results pointed out that 30 percent of the country’s top 50 websites made use of **<input type="email">** as their solution for the validation of e-mail addresses on the browser side of the equation. The team expected to find a similar pattern in the global survey, and that proved correct. The rate of global use of this string ranges somewhere between 20-30 percent according to the patterns we have observed.

The problem is that in mid-2019 the HTML5 standard has not yet caught up with Universal Acceptance. This leaves us concerned as deployment of the standard continues to grow and developers increasingly rely on its functionalities to get their code working under as many scenarios as the current state of the Web demands.

This is the current pattern of acceptance for **<input type="email">**:

| Test case | Result |
|---|---|
| ascii@ascii.newshort | Accepted |
| ascii@ascii.newlong | Accepted |
| ascii@idn.ascii | Accepted |
| unicode@ascii.ascii | Rejected |
| unicode@idn.idn | Rejected |
| arabic.arabic@arabic (RTL) | Rejected |

This result only reinforces our previous assertion that the key misconception seems to be that symbols should not exist before the @ character.

If 20-30 percent of the websites make use of this string, then conversely, this means that the same number of websites is failing this stage of validation due to a lack of proper standards in HTML5. In other words, as far the browser side of the equation is concerned, this is the most significant action that can be taken to further UA at the moment.

Having investigated the status of this issue, we found that the 5.3 draft of the specification, on the section related to e-mail validation[6], predicts compliance with RFC 6531 and RFC 5890, making the standard compliant with UA. However, this revision has been in “first draft” status since 2017 with no predicted date for it to become a “candidate recommendation” or a final “recommendation”.

We find this concerning, given the importance of the matter in order for the community to reach Universal Acceptance, particularly as it relates to internationalized names. This is an issue the community cannot let sit idle, and broad steps need to be taken in our relationship with the W3C, WHATWG, and browser vendors, in order to speed up the process of implementation for this aspect of the specification.

---

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

This is particularly timely as the W3C and WHATWG have recently agreed on taking a joint approach towards standards development[7]. Reaching out to relevant parties and demanding that an “eaimail” option be added for websites that are UA-ready is important, just as much as pressuring major browser vendors to signal that they are interested in making this option viable.

The research team recommends that the resolution of this issue is made into a high priority.

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

## Can These Reports Be Automated?

During the elaboration of this study, questions were brought forward by the UASG in ICANN meetings and during planning sessions about the feasibility of an automated software for the production of reports such as this one. The team tried to find a practical answer to that question but arrived at the conclusion that no, it is not possible to automate these tests at the moment.

These are our main concerns:

- There is too much variety in terms of how websites are coded, what scripts they are using, and how the forms are organized. Overall, it would be practically impossible to account for all variants.
- Minification is widely employed in order to reduce loading times of websites and scripts. At times, this requires somebody with coding experience to look over the results.
- A significant number of websites require some sort of Captcha or have some nuance to their forms, which renders automation useless.

Some of the same rationales can be applied in the question of the outsourcing of the study to non-experts such as in the case of using Amazon's Mechanical Turk. While that may work in terms of performing the actual testing, coordination would still be necessary to prepare the list of domains, extract the code, and then produce a meaningful report that's more than a statistic. Still, it is much more viable than automation.

## Conclusion

If we take into consideration the top 100 websites in the world as a reference, only five of them accepted all test cases: “quora.com”, “espn.com”, “spotify.com”, “txxx.com” [NSFW], and “godaddy.com”. Suppose we were to skip the difficult “unicode@idn.idn” and Arabic (RTL) test cases, we would still end up with only nine websites accepting the first four test cases. Since we are talking about the most accessed pages in the world, this is quite concerning.

While we are optimistic about the prospect of newshorts reaching full compliance in the near future and find it realistic that newlongs are headed in the same direction, it is clear that the internationalized domains need a boost and that proactive action needs to be undertaken in order for them to thrive. It is not only that stronger policies need to be developed, but rather that the Internet needs to be made more aware of the existence of such e-mails and domains – something that will only be achieved by proactive outreach.

In conclusion, we believe that an important next step would be the development of a dashboard that would allow for the evaluation of all reports on UA that are generated employing the metadata schema proposed in this document. This would generate a long-lasting repository allowing for realistic measurements to be made by any interested party and would serve as a reference tool for the entire community.