# UA Measurement WG Meeting

03 November 2022

**Attendees**

Nabil Benamar
Harsha Wijayawardhana
Jim DeLaHunt
Shreedeep Rayamajhi
Sushanta Sinha
Mallan Gouda
Julien Bernard
Arnt Gulbrandsen
Yin May Oo
Seda Akbulut

**Meeting Agenda:**

1) Welcome and Roll Call
2) Continue discussion on M4 Action item: Characterize how much Android platform limits acceptance of IDNs in web browsing (UASG037)
    a. Identify the UA related constraints on applications running on Android platform)
    b. How does UTS #46 differ from IDNA 2008?
    c. Define the outcomes and objectives of the work

**Meeting recording**

**Meeting Notes**

In this meeting, only the M4 action item topic was continued to be discussed more in detail. Seda introduced Arnt to the WG as the new UA Technical Senior Manager at ICANN. Everyone welcomed him.

Jim referred to his own notes, and recap what was discussed in the last meeting. We discussed the document (UTS#46) listing specific words in specific languages, domain names involved with those listed “four characters” would matter a lot. The focus was on characters of deviation such as German letter Eszett, where Harsha gave a lot of information on the difference of having and not having the zero-width joiner (ZWJ) is very significant in Sinhalese script. **It would be useful to have a document listing language specific details about name in which would differ between IDNA2003, IDNA2008 and UTS#46.**

Julien suggested l**isting the domains under gTLD to see how many domains would be impacted.**

**Arnt said he was already working on this.** He was waiting to get the access now. But we will not know how many 3rd level domains are there. People can register anything at the third level. Jim agreed that it is a good idea to have a list of domains and wondered if there is really a way to get the list of SLDs?

Jim said that the rules are different for top-level domain names, 2nd-level domain names, and third- and lower-level domain names. For example, are there any .com domain names with a ZWJ in them?

Arnt said some countries do not allow sharing domain names, however, some registries provide some sort of list to ICANN as confidential data, just to report. And we know that there are good collaborators for us to work on some issues if we were to make a list of a complete list of domains and trademarks that are not registered a domain yet. Jim said it is not to know what these domain name are, but to know if there are any domain name which uses any of these listed four characters in the UTS#46 is sufficient for this purpose.

Related to the okHTTP report contacted with UTS46 maintainer, Arnt mentioned that he knows that person. There is a difference between IDNA2003 and 2008. However, in the real world there is no difference for domain names. **There are no domains in the world for which this matters.** In theory, we can treat this as not an urgent issue. Otherwise, this will be a time-consuming task for us.

Harsha questioned the disallowing of zero-width character rules, ZWJ and ZWNJ, when GP implemented the LGR rules unaware of the IDNA2008 specs. Currently for Sinhala script, they are allowed at the 3rd level, but not at the 2nd level and TLD. Arnt suggested contacting the UTS#46 maintainer, however, Harsha rejected that it is not important to change. Although TLD completely blocks zero-width characters, the second level and lower might have issues with it.

Jim mentioned that in many discussions, people got lost talking about the theoretical impact of software following IDNA2003 vs IDNA2008, without being able to quantify the affected domain names or scripts or which language group.

**Having the data would be helpful to understand its impact.** Having people who know those affected languages or scripts would be really helpful as well, having UASG to write down which languages are impacted would be important.

Harsha explained that during the Unicode implementation, he defended that those letters needed the zero-width joiner to be encoded. But the final decision of the Sinhala GP was to block them out. We are looking at possibilities. Myanmar did a similar thing as well. We need to be very careful. We will have some issues with Sri at the 2nd level. After IDN2003, there was a long debate where it was allowed in the IDNA2008. But it did not correct the issue.

Jim asked **who could write this issue in English to document to explain the impact for people who do not know the script.** Harsha said he already had a similar document and he can develop a writing around this issue.

Jim looked back to the "okHttp-bug-report" issue where they replied the **difference was not important.** Jim said we need to put resources to measure and remediate to see if it is Important or not.

Harsha pointed out to this paragraph of the document:
https://unicode.org/reports/tr46/#Table_Deviation_Characters
"Because of the visual confusability introduced by the joiner characters, IDNA2008 provides a special category for them called CONTEXTJ, and only permits CONTEXTJ characters in limited contexts: certain sequences of Arabic or Indic characters. However, applications that perform IDNA2008 lookup are not required to check for these contexts, **so overall security is dependent on registries having correct implementations.** Moreover, the IDNA2008 context restrictions do not catch most cases where distinct domain names have visually confusable appearances because of ZWJ and ZWNJ."

Yin May suggested in the chat for Harsha to take a look at Myanmar Script Rootzone LGR for handling codepoint sequences since Myanmar script and Sinhala script have similar background, although Myanmar script handled rendering differently without using any zero-width character. This might help, however, Sinhalese GP should discuss this separately if there is a need to amend the LGR rules. (Myanmar script do not use zwj though)

Seda highlighted the example under the same session when the IDN domains in .de TLD, have a URL with the double "s", which could be mapped to another label with German Eszett character.

"Alice's browser supports IDNA2003. Under those rules, http://www.sparkasse-gießen.de is mapped to http://www.sparkasse-giessen.de, which leads to a site with the IP address **01.23.45.67**."

Seda asked how they would handle this security issue (there must be some rules for this), and suggested Sinhala GP might want to look at those as well.

Arnt added to Seda's point that he once heard when IDN domains were introduced in .de, there was indeed such a rule. Seda shared the current guideline https://www.denic.de/en/domains/de-domains/domain-guidelines/

Jim thanks Julian for checking, and then explained that UTS#46 said this is security risk because if you go to a domain name involving an Eszett, IDNA2008 will redirect to a different site than IDNA2003 browser, and the site which is supposed to be supported by IDNA2008 browser is currently not registered yet, actually ".de" wouldn't allow this for security reasons.

Jilien shared his screen to show how to go to the website with A-label form in the **FireFox browser.** Jim thinks there is a problem with either Chrome or Firefox browser.

Arnt explains that **".de" can only allow the same owner to register both variant labels.** German language users use the double s and the Eszett interchangeably. Variant labels being owned by different parties is where the security thread lies. So, the variants of the same origin should be mapped to the one same website.

Jim wants to discuss how big would this problem be (the situation of Eszett being converted to ss) in IDNA2008 if our overall goal is Universal Acceptance. We should observe more what the main obstacle could be.

Harsha confirmed his action item.

Arnt quickly commented that **changing from IDNA2003 to IDNA2008 should not be an obstacle.** He tried to find a big example but found only in Ethiopic, Yiddish and Emoji domain names. The bug-reply from okHttp was an unfortunate point.

The discussion generally gave the impression that the difference between IDNA2003 and IDNA2008 was not a major problem, as the difference between IDNA2003 and IDNA2008 was only 4 characters, some of which were already blocked by registrars, and some registrars may have different security measures, such as registering variant domain names to the same registrant.

Jim suggested talking about “**Is IDNA2003 vs IDNA2008 a big deal?”** the next time. If this is not an important thing, we should **look for a more important objective.**

**Next meeting:** Thursday 17 November 2022 UTC 1600-1700

### Action items

| No. | Action Item | Owner |
|---|---|---|
| 1 | Share the document of the impact of having ZWJ in Sinhala script domain names | Harsha |
| 2 | Write down which languages are impacted by the listed four characters, comparison of allowing and disallowing these characters. | Measurement WG |