# UA Measurement WG Meeting

01 December 2022

**Attendees**

Harsha Wijayawardhana
Jim DeLaHunt
Sushanta Sinha
Julien Bernard
Carine Malor
Frank Anati
Arnt Gulbrandsen
Yin May Oo
Seda Akbulut

**Meeting Agenda:**

1. Welcome and Roll Call

2. **What's next with the M4 Action item?**
   **Is IDNA2003 vs IDNA2008 a big deal? or should we look for a more important objective?**

   M4: Characterize how much Android platform limits acceptance of IDNs in web browsing (UASG037)
   .Identify the UA related constraints on applications running on Android platform)
   i. How does UTS #46 differ from IDNA 2008?
   ii. Define the outcomes and objectives of the work
3. AOB

**Meeting recording (1 Dec 2022)**
https://icann.zoom.us/rec/share/lEYxRcaXoe8ENDtEp2VeZ7IuKLhb4U8X8JPt1bHjIIZip2bwT2S_zCovXkXtGXVN.oZ_AIcJhmJC7ufNX

**Meeting Notes**

Seda recapped agenda items and leaved the floor to Arnt.

Arnt said he has got necessary files to gather data of real world domains and these would be prepared for the next meeting. We can talk about this when we have more details.

Jim shared something he learned from a recent EAI working group meeting, which was about email interfaces displaying domain names in punycode. Jim assumed that changing the unicode name to punycode might require IDNA2003 or IDNA2008 processing. On this topic, Harsha, a Sinhalese native speaker had discussed the importance of IDNA2008 and the difference between the choice, and IDNA2003. According to him, the invisible character Zero with joiner (ZWJ) is used a lot in the names of Sinhalese. We need to be able to answer to those script users who are affected by IDNA2003. Since we do not live in an ideal world, we could not enforce IDNA2008.However, we should be able to document the significance of differences and the impact of using IDNA2008. He shared that the right thing for the UA community is to work as hard as possible to have/help everyone adapt to IDNA2008. **So we should have a resource that will allow the UASG community to learn about it more**. Jim referred to Marc Blanchet's email back in August about pushing IDNA 2008 in the mailing list.

Jullien said we could not accept two standards at the same time, it could lead to compatibility and security issues. In any case, we only had one choice, IDNA2008.

Arnt said software would always have bugs, if they used IDNA2003, what they would do is to invest writing codes in telemetry to fix any issues. It might seem too hard from their point of view to adapt to new implementations when there are any changes**. Our shared code samples should make the implementation of IDNA2008 seem as easy as possible. It would be better to share more of IDN and UA compliant codes even if they have bugs, it would always be better to have buggy codes than no codes.** Jullien agreed and said allowing people to add more codes would help. One point was the standard libraries of java and python were still at IDNA2003, this could be changed to IDNA2008 as default and then the problem would be easier to be solved.

Jim pointed out that there is an interlocking behavior of software which process domain names and the people who operate websites and choose domain names and domain registrars who set policies on what names are allowed, and users who try to use domain names. This is not purely a matter of making an opinion on how a software should be written. This is trying to chart a strategy to keep all pieces moving forward to a destination while being careful to avoid situations at which point the software did not support IDNA2008, and then, a whole bunch of

domain names were picked and that became a defacto standard for domain names and became an obstacle to move to IDNA2008 in the future. In short, this is more complicated than software compliance.

Arnt said that this is the main reason he is working on zone files. To find a problem in reality. The number of existing domain names which would be affected by the change would be small. For example, the Android library that is moving towards IDNA2008 will not cause any difference in behavior in practice, so it is a safe thing to do so.

Harsha expressed his concern about not being able to use the Sinhala script correctly if IDNA2008 was disallowed. Arnt assured Harsha that this was why he was pitching to see more IDNA2008 compliant codes to be shared, because fixing bugs is easier than implementing the whole code for a feature. Harsha said he checked four browsers at the moment and only Firefox is IDNA2008 compliant out of 4 browsers, the others are Opera, Microsoft Edge, and Chrome.

Jim also said that current Sinhalese TLD LGR do not allow the ZWJ that is required for Sinhalese words and allowed in IDNA2008; therefore, he assumed there is no negative effect on Sinhalese domain names when browsers change from IDNA2003 to IDNA2008. Harsha agreed, and emphasized that when IDNA2008 is adapted more widely, Sinhalese panel would allow the ZWJ for the second level and onwards.

Jim shared that Sinhalese code writers are the ones that are more impacted by this difference in specification more than German code writers.

**Jim requested Harsha to share his observations through a presentation or a document. Harsha said he would share within a week.** Julien supported the idea that the opinion from someone who knows that language is very important.

Harsha said the Sinhala LGR Generation Panel has started to work on allowing the ZWJ for the second level domain names. Seda suggested that having supportive examples of how ZWJ helps forming the characters would help non-Sinhalese speakers to understand the challenge. Jim said it better highlights **how destructive it is to disallow ZWJ in Sinhalese script with examples or comparisons**.

**Examples in the chat:**

(1) imagine that one was not allowed to use “L” or “O” in English text. One must use “1” and “0” instead. The company “Global Logistics” would have to use a domain name “g10ba1-10g1st1cs.c0m”. That is very difficult to read, and looks ridiculous in English.

(2) imagine that one was not allowed to use “w” in English text. One must use “uu” instead. “Waste World” would have the domain name “Uuaste Uuorld”.

(3) imagine there is a ZWJ requirement for latin ‘æ’ ( a + zwj + e = æ )

(This may not be the case for Latin script characters, since Latin combined characters have their own unicode codepoint, however, this happens a lot in complex script languages.)

Julien said reading the French work oeuf vs. the correct for œuf is not that bothering so it will certainly explain the problem but may not reflect as bothering it can be for some languages.

YinMay suggested waiting for Harsha input as native speaker would help the most.

Harsha said he initially thought there would be problems with Devanagari scripts, and then, he figured those scripts required ZWNJ while the default form is always conjunct, and for the domain names, it is fine with the conjunct forms. Jim said having examples which would be easier to understand for European, Latin American people to understand would help a lot while they do not understand complex-scripts. Jim said that a big part of this work is educating the people and he suggested that **it would be great to have examples for all four languages featured in the four-characters table, and how they are affected by IDNA2008 vs IDNA2008.**

Harsha explained that because of the typewriters issue in the past, the broken form which does not require ZWJ was commonly used in Sinhalese printings before the 20th century. Now the Sinhalese native speakers would like to use the historically-correct conjunct form which requires ZWJ. During the time of implementing the Sinhalese unicode, ZWJ was suggested to use when Sanskrit-derived words are required. Nowadays, these words have become radio station names or business names and have potential to be used as domain names.

**Harsha said he would share his observations, and in fact he wrote an article along with statistics on this and will share it with the working group mailing list tomorrow.**

Jim said **Harsha could ask through the mailing list to let other language users know if there are any script users who have knowledge of ZWJ as well.** Jim said it was only our assumption that Devanagari is not affected by having or not having ZWNJ, however, Arabic is affected for sure, so we need more input from native speakers of other script languages.

It would be valuable for us to describe that so that people from other parts of the world can learn that. The difference between IDNA2003 and IDNA2008 becomes an issue. It was thought how to do metaphors for English language and Latin script to show the damage you get from not being able to use all characters, so it would be valuable for us to put that in writing in a form where people from other parts of the world can understand the impacts.

Julien appreciated Harsha's effort on the article. He asked whether Harsha is aware of any bug reports with the Chrome team. Often Germans are asking for issues about essets, but he doesn't recall bug reports for Sinhala. If people complain about those kinds of things, they will finally implement it, also for the edge.

Arnt shared that the Greek Sigma problem was prevented by the Greek LGR Generation Panel, so there is no issue. Therefore, out of 4 characters allowed by IDNA2008, the Greek Sigma could be left out to be investigated. **So the difference between IDNA2003 and IDNA2008 is not harmful.** Basically it is only 3 characters, and not 4.

Harsha said that the TLD was delegated in 2010. It's been a long time since implementation. Not many people adopted these ccTLDs. He realized that nowadays there's a big demand on these two Sinhala ccTLDs. So they need to immediately have a variant process.

Jim stressed the importance of documenting this. We have to explain why it matters. And show how big a problem this is.

Jim shared in the chat:
-I suspect that the differences boils down to what UTS #46 calls the four "Deviation Characters": U+00DF Latin Letter Esszett ß, U+03C2 Greek letter sigma

ς, U+200D ZWJ, U+200C ZWNJ. We have people in UASG who know many languages. It would be good to hear from them about other languages affected by these IDNA2003 vs IDNA2008 and these “Deviation Characters”.
Jim added that UTS46 gives an example of ZWNJ for Arabic.
https://unicode.org/reports/tr46/

**Next meeting:** Thursday 15 December 2022 UTC 1600-1700

**Action items**

| No. | Action Item | Owner |
|---|---|---|
| 1 | Share an article along with the statistics with the working group mailing list | Harsha |
| 2 | Share some results about the zone files in the next meeting | Arnt |