Discussion of Possible Ways Forward

To be led by BoF co-chairs

Dave Thaler & Marc Blanchet

Some questions

1) Do we agree there is an important problem?
2) Do we agree on what the problem(s) are?
3) Is it possible to mitigate at least part of the problem(s) or is it hopeless?
4) Is there a direction that the IETF and Unicode Consortium can do complementary work on?
5) Is there anything the IETF should do by itself?

Problem(s) [TO BE EDITED LIVE DURING BOF]

- Multiple abstract characters with same glyph+script+properties in identifiers without language/locale info, inherently allows user misunderstanding
- …

Directions

This is only allowed if we think we ought to solve stuff!

1. Find them, disallow new, cope with old
2. Disallow some combining sequences
3. Just warn
4. Get a (or >1?) new Unicode property
5. Create NFI(s)
6. Bundling/blocking

Backup

(not part of presentation)

Which steps(s) have the problem(s)?

- IDNA:
  - Input -> [RFC5895 mapping] -> [NFC] -> [Comparison]
  - Input -> [RFC5895 mapping] -> … -> [Display]

- More generally:
  - Input -> [Mapping] -> [Normalization] -> [Comparison]
  - Input -> [Mapping] -> … -> [Display]

draft-klensin-idna-5892upd-unicode70-04 suggests property(s) might need to provide

1. Identification of combining characters that, when used in combining sequences, do not produce decomposable characters.

   [[CREF2: Wording on the above is not quite right but, for the present, maybe the intent is clear.]]
2. Identification of precomposed characters that might reasonably be expected to decompose, but that do not.
3. Identification of character forms that are distinct only because of language or phonetic distinctions within a script.
4. Identification of scripts for which precomposed forms are strongly preferred and combining sequences should either be viewed as temporary mechanisms until precomposed characters are assigned or banned entirely.
5. Identification of code points that represent symbols for specific, non-language, purposes even if identified as letters or numerals by their General Property (see Section 3.3.2.2 and Section 3.3.2.1).

So what's the problem?

This is what caused the realization:

So, just combining marks? No!

- க (U+0B95, Ka) vs ௧ (U+0BE7, digit 1)
- ٣ (U+0663, digit 3) vs. ۳ (U+06F3, digit 3)
- 口 (U+53E3, “mouth, gate”) vs. 囗 (U+56D7, “proud, upright)