Public Comment to Expression of Interest: IDN briefing

1. The Unicode© project started in 1991. The Common Locale Data Repository (her in after CLDR) project started in 2003. The public available Internationalized Domain Name (here in after IDN) registration started in 2003. The top level roots started serving as IDN 2003 based on Unicode 3.2. The IDN 2003 means a couple of IETF’s Requests of Comments (here in after RFC) 3454, and Proposed Standards numbered as RFC 3490 RFC 3491 and RFC 3492 (and a very special Informal RFC 5242.) The RFC 3743 suggest far eastern language regulations. The RFC 4290 and RFC 4690 are suggesting some extensions for IDN to dissolving early time restricting rules for filtering. Many of IDN related RFCs and Unicode Technical Standards studying from the viewpoint of confusing and of misleading Unicode elements resulting homograph attacks. These early lists of resources restricted a lot of Unicode© code-points from IDN registries. The CLDR and the set of Unicode Latin range listing not matching of spoken alphabets from Latin range. The Unicode code point assignment focuses to accent like modifiers of the elements based on English alphabet. The 5th chapter of RFC 4690 suggests forwarding practice for IDN improvement. While the Unicode system not allows the revoking assigned code points, the special timed RFC 5242 could turns from joke to reality: there are a couple of languages representing by Latin scripts, where tailoring is distinguishing. The Locale Explorer of CLDR lists main resources for Locales and linguistic data. The CLDR is not checked and not informal what Latin scripting languages having sequences, where the tailoring – means no hyphen inside allowed on said sequence – is restricted. The other missing resource of CLDR is the mapping of the Latin Alphabet of Locales to Braille cells. Generally this time the six dots Braille cells, representing 64 code values, are in use. Some of Latin based Braille scripts are using non-tailoring rules for letter like expressions of sensing data outputs. The normal number implementation of Braille uses the Number follows sign generally represented as 3456 Braille cell. The Braille number implementation stops with space. Different Latin scripting Braille systems have no sign sets for Euro Currency, for COMMERCIAL AT and for many other useful sings of modern computing. The classic six dot Braille cells have no assignment for language distinguishing. For the easy recognition of modern computerized Braille some standardization with extended Braille cells as x78 Unicode© code points could be helpful. Most important extensions are:

NEW LOCALE meaning: 28FF as DOTS-12345678 as language neutral;

NEW NATIONAL LANGUAGE SYSTEM meaning 28FE as DOTS-2345678 as language neutral;

0040 the COMMERCIAL AT 28C7 as DOTS-12378 as language neutral;

20A0 EURO CURRENCY 28CE as DOTS-23478 as language neutral,

20AC the EURO CURRENCY 28D5 as DOTS-13578 as language neutral.

It is not only sensibility extension on typography but allows separation of Latin spoken languages for visually impaired persons.

2. The continent wide registry of Europe, that is Eurid started as internationalized domain name registry according to 733/2002 EC Regulation. The 5th Article (d) regulates languages and geographical concepts. The Eurid started when the 874/2004 EC Regulations had been forced. The 560/2009 EC Regulation amends some parts of 874/2004. The language regulation of first, second, third and fourth paragraph of 6th Article of 874/2004 is amended as: “If it becomes technically possible to register names in the official languages under the .eu TLD using alphabetic characters which were not available for registration at the beginning of the phased registration period provided for in Chapter Ⅳ the registry shall announce on its website that it will be possible to register names under the .eu TLD containing those characters.” (The tld is the abbreviation of top level domain. The Eu tld is categorized to ccTLD.) The Unicode CLDR not lists Locale and linguistic resources for European Community. Now the Eurid restricts the use of ß, but not propagates the linguistically correct and assigned Unicode© values. When the Eurid started its own phased registration, it was not propagated characteristic elements of Eastern European languages, and not checked the correct elements –the spelling – of its own Whois. The Eurid started its own registry from one of the early started IDN top level domain root. That root where Eurid started has official Hungarian resources same as IANA listed. That time the German ß had been transformed to ss as tailoring ss structure, but restricted the PatentPendingCOMMERCIAL(AT)**xn--Rswk-vzac.EU** Punycoding-decoding**.** It is not clear what is the technical background and standard that runs under Eurid.

That background is surprising if we check the first sub variant of first Unicode release, where we can find those resources what are abandoned by from the starting of Eurid. A four year operation is not enough for that ccTLD registry to revise and correct its bugs from starting.

BestPractice is keeping bugs and combining with mixed scripts:

Fig.1.:



Fig. 1. is the CLDR Locale data for Hungary. Exemplar Characters are not corrected. Earlier version had a bug for GQ tag now it is corrected. Most of di-glyphs having Braille cell mappings.

Fig. 2. :



On Fig.2.the blue text is newer and consist of correct Hungarian element and black is incorrect.

Smart TLDs maintaining filtered IDN Registry are distinguishing and rejecting the highlighted black text and accepting only highlighted blue variant. Not to forget reading Whois Policy and Terms and Conditions.

Fig.3.



On Fig.3.the correct Hungarian encoding of 0151. If it is a variant with 00F, like illustrated on Fig.4 it is not acceptable for Hungarian IP range:



On Fig. 4 the linguistic data is false; the 00F5 is restricted by NIC.hu (Domain.hu)

Fig.5.



On Fig. 5 is the only one Hungarian di-glyph having no Braille cell assignment. That di-glyph is member of Unicode 1.1 since 2001, and member of Hungarian, Polish and Slovak alphabets. Suggested Braille cell is 28CF DOTS-123478. The xn--vka is different from dz ccTLD representation of Algeria.

The Eurid started its operation without running IDN service and maintaining correct linguistic rules from an unknown reason.