Subject: Re: White paper from Dr. Raymond Doctor
Date: Sunday, August 14, 2011 4:57:49 AM PT
From: doc
To: Andrew Sullivan
CC: Naela Sarras, Steve Sheng, Francisco Arias, Dennis Jennings, Nicholas Ostler, akshatj@cdac.in, nehag@cdac.in

Dear Andrew and colleagues,

Please allow me to thank Andrew Sullivan (and Nicholas Ostler: thanked in an earlier post) for taking time off to read the white paper and giving it so much attention and care and providing such thoughtful comments. I know how busy they are and am all the more grateful for their comments and suggestions.

In fact the two sets of comments complement one another. While Andrew Sullivan's examines the technical issues, Nicholas Ostler looks at the paper from the Linguistic point of view. In spite of a few overlaps (just 2) their comments and suggestions are thought-provoking and extremely useful.

I am posting a reply to Andrew Sullivan's comments in an attachment, not as a defense but more as an attempt to understand better my motivations in writing this white paper..

Thank you once again for all your kind help and suggestions which I trust will be helpful to all who read them.

With all kind regards

Doc

On Thu, Aug 11, 2011 at 11:40 PM, Andrew Sullivan <ajs@anvilwalrusden.com> wrote:

> Dear colleagues,
>
> On Wed, Aug 10, 2011 at 11:52:43AM -0700, Naela Sarras wrote:
>
> > The variant project team members included in this message are kindly
> > requested to review the document provide any comments, critique, or
> > suggestions to Dr. Raymond Doctor. Specifically, I am asking Andrew
> > Sullivan and Nicholas Ostler to review the document and give
> > feedback . Please provide your feedback by close of business (PST
> > time) on Monday, 15 August 2011. Once the project team has provided
> > their feedback to this white paper, the document will be shared on
> > the vip@icann.org mailing list.
>
> I have read this document. I thank Dr Raymond Doctor for preparing
> it, and for the Devanagari team for providing it for review. It is a
> thoughtful and broad response to the initial questions, and I believe
> it is extremely useful for the purposes of this project. I have some
> detailed comments, which I lay out below. I have keyed my comments to
> the section numbers in the paper.
>
> 2.1, Abstract Character.
>
> This section argues that the term Abstract Character can be
> understood as meaning the same as Glyph. But the Unicode Standard
> explicitly denies that these are the same:
>
> - An abstract character has no concrete form and should not be

confused with a glyph.

- An abstract character does not necessarily correspond to what a user thinks of as a “character” and should not be confused with a grapheme. (see Unicode Standard 6.0, 3.4 D 7, p 66)

Since one goal is to cleave to external sources for definitions, I think we must say here that Glyph and Abstract Character are not the same. Moreover, the example here is in terms of the bilabial unvoiced stop /p/. I think that it is important to recognize that there is a siginficant link between a language and its writing system(s), but I am under the impression that the Unicode Technical Committee (henceforth, UTC) tries very hard to stay out of making linguistic decisions and to stick on the side of encoding writing systems. I suspect, therefore, that an abstract character is more like "the thing that looks like A with a round thingy over the top", which happens to be encoded in Unicode as U+00C5 LATIN CAPITAL LETTER A WITH RING ABOVE, U+0041 LATIN CAPITAL LETTER A plus U+030A COMBINING RING ABOVE, and U+212B ANGSTROM SIGN.

2.1, Language Character Repertoire

I think the discussion here is entirely congruent with the reason the JET guidelines originally conceived of variants in terms of a language rather than a script. But there is a difficult problem raised as a result, which is that in the context of zone policy, one has to make choices that resolve registration request conflicts in favour of one language or another. In the root zone (as well as in gTLDs), such a policy could be problematic, which is why there has been considerable effort to reduce issues to a script rather than a language. I'm not sure what to do about this sort of problem: any inter-language problems are going to be extremely difficult to address in these kinds of zones.

2.1, Summing-up [some discussion of normalization]

The discussion here suggests that it is important to discuss normalization and to test browsers to check that they actually perform such normalization. This issue has also come up in the Arabic VIP. I think some important distinctions are needed.

First, we need to ascertain whether we are talking about Unicode normalization or something else. Unicode normalization has four forms: NFC, NFKC, NFD, and NFKD. I could explain all of that here, except that I would just repeat what is in http://www.unicode.org/reports/tr15/tr15-33.html. If you haven't read and understood that, however, none of what I am about to say will make any sense.

There remain, as I understrand it, some kinds of normalization that are not actually covered by Unicode normalization. This normalization is generally linguistically sensitive. So, for

instance, in some Arabic-script using writing, there are modifier dots that have no effect on the meaning of a piece of text, and can be written or not as a writer likes. (But I am told that they are not handled by the relevant Unicode normalization in this case. More below.) I was under the impression, after the Pune meeting, that no Devanagari-using language has this issue, but as I speak none I'm hardly going to be useful in coming up with counterexamples.

Now, as for the normalization we need, we have two kinds that affect us. IDNA2003 -- specifically, the Nameprep step -- uses NFKC. This is not optional, but the use of compatibility equivalents was one of the things that people didn't like in IDNA2003 (because it meant that you could not be sure that the transformation to Punycode and back again didn't lose any data: things like final form sigma would get mapped away).

IDNA2008 solves this in two ways. First, it does not itself do any mapping. However, it defines U-label such that only things in NFC form can be U-labels. How the string in question gets into NFC is not part of the protocol, and is something that is supposed to happen before IDNA2008 takes over. Second, IDNA2008 disallows any character that is not stable when performing NFKC and caseFold (the actual rule is B: toNFKC(toCaseFold(toNFKC(cp))) != cp in RFC 5892 section 2.2. See http://tools.ietf.org/html/rfc5892#section-2.2). The goal is, roughly, to make sure that only stable characters are candidates for being used in the protocol, but otherwise to stick to NFC.

It might seem tempting, then, to try to ensure that applications are using just the right normalization in all cases, and to try to add policy where it is clear that some clients are not following the protocol. There are, however, two problems with this approach.

The first problem is that, if a client isn't using NFC for its U-labels, then it just isn't implementing IDNA2008. Similarly, if a client isn't performing NFKC on its strings, then it just isn't implementing IDNA2003. It is probably wise to have registration policies with some sort of sunset clause that results in the most restrictive policy for the intersection of these two protocols (effectively, one wants a policy that recognizes that IDNA2003 is still more widely deployed); but that is not the same as saying that one is going to test for bugs and implement around them. Testing for nonconforming client software on the Internet is shooting fish in a barrel, and if you try to make policy around the bugs in Internet client software, you will quickly go mad. But more importantly, at least some clients can accept and use raw UTF-8 as part of their DNS lookups, and that is regarded as a feature rather than a bug by the users. Such clients aren't implementing IDNA at all, but are using UTF-8 in the DNS. This is perfectly legal: DNS labels are not LDH-labels, but bits. It is only by convention (really, really widely held and widely regarded as protocol, but still a convention) that DNS labels are

restricted to letters, digits, and hyphen.

2.3.2.1 Orthographic alternanats

While I like the term "alternants" as a term we might want to adopt, I am having a hard time understanding why the examples we see in this section do not lead us directly to say, "Color and colour are alternants, and by alternant policy the registrant of .color must also received .colour too." If there _is_ no principled reason why not, then why isn't the converse true. That is, if the case in English works out the same, what is so special about the non-English cases that the answer should not be, "If you want two labels, register (and, I presume, pay for) two labels"?

2.3, Table 4

First, I should note that I think this table is very helpful. I appreciate its development.

In the first row of this table, item 2, there is the observation that new versions of Unicode will cause problems. This is true, but it is true _forever_ under IDNA2008. That was part of the point. Registries will always need to have sunrise provisions to cope with the fact that new characters will show up in later Unicode versions. A policy, of course, could be, "We won't register them," but then you still have the problem of code points that are PVALID under an earlier version of Unicode, but become DISALLOWED under later versions.

This is a feature, and not a bug, of IDNA2008. IDNA2003 was pegged to a particular version of Unicode, but that didn't work. When you upgrade your computer and get a new version of Unicode, you can't tell which version of Unicode you're using. So many of the putative IDNA2003-implementing clients out there aren't actually implementing any standardized protocol at all, because they have a later version of Unicode and can't tell. The only way to solve this issue is to forbid UTC from changing Unicode, and we can't do that.

A later row of the table discusses case marking. I think this is not a problem for anyone, even though some think it is. IDNA2008 is designed specifically so that upper-case letters aren't allowed in U-labels. It's true that this is frustrating when compared with the special processing of A-Z. There's nothing that can be done about that special processing, however, because it's a built in part of the DNS protocol.

The last line of this table is potentially useful input to a specific policy, but I do not know how any general policy can be made on the basis of browser issues. This is for two reasons. First, there is absolutely nothing that we can do about different browsers on the Internet. We cannot detect what browsers people

are using when they do DNS lookups, so we cannot react usefully to such differences. Second, it is not only browsers that are relevant. DNS names are used all over the place: in mail agents, in SIP VOIP clients, in system configuration files that only system administrators ever see, in system logs, and in machine-to-machine transactions in which no human is directly involved most of the time.

2.3.3, Problem of the Preferred Variant

In this section we have a proposal for a way of identifying a number of alternants for one another, and then picking which of those is to be allocated and delegated.

Suppose we have a set of alternants, {a1..an}. These are each expressions of the Archivariant. The central proposal of this section is that, on registration request for any one of {a1..an}, the registry is checked for conflicting members of any other set. If there is no conflict, the entire set {a1..an} is allocated to the registrant. This allocation does not, however, automatically result in delegation. At this point, it is a matter of registry policy how many of the individual alternants, if any, are to be registered; and what that will cost.

I would like, however, to distinguish a couple of terms here that I think are used in more than one way. I want to say that "bundling" is not what has happened in the case of www.color.com and www.colour.com: those are actually just separate registrations that happen to resolve to the same IP address (note that they don't go to "the same page". One of them is a redirect to the other, and if you visit http://www.colour.com you will be redirected to http://www.color.com. This is an http-level redirect and not something done in the DNS). I think it would be better to use the term "bundling" for the case where the registry automatically performs the linking of the different registrations together somehow, as happens here when the allocation of all members of {a1..an} go to the same registrant.

By the same token, I think it would be good to distinguish reserved and blocked names. We might want to say that a reserved name is not allocated at all. All two-character domains in the top level that are not actually registered are in fact reserved: nobody is allowed to register them until ISO allocates a country code. But a name that is allocated but not delegated is blocked: nobody else can register the name because, in effect, it is already registered.

I hope these comments are useful.

Best regards,

Andrew

--

Andrew Sullivan
ajs@anvilwalrusden.com

--
Best regards,

Doc