IDNA2008 Situation (DRAFT 1.0 2/1/2010)

When IDNA2003 was developed, it introduced the idea of mapping some Unicode characters into others. The argument was made then that having a commonly adopted mapping scheme was a good thing.

A different view is that mapping is not a good idea. This view arises in part from a strong sense that the notion of a “canonical” domain label form is an important contributor to stability for internationalized domain name.

Some mapping may be “needed” to cope with local situations. For example, it might be more natural to express domain name label separation with a character other than “period” (U+002E) because the local user’s keyboard does not easily allow input of that character. However, such local mapping could (should?) be confined to local input transformation and not allowed to alter what I will call “canonical representation” of a domain name. If the substitute “dot-oids” escape from purely user interface artifacts into domain name representations presented in other than the originating local context, then applications that know nothing about IDNs may (will!?) be unable to parse the domain names correctly, even if the labels themselves are in ACE (punycoded) form. It seems imperative to assure that only U+002E is ever exhibited as a valid label separator except at the user interface edge where it should be translated immediately into canonical form.

IDNA2008 deals with labels. It defines as canonical the label separator “period”. This suggests that for display and absolutely for “on the wire” exchange, or storage in files that might be processed by non-IDNA-aware applications, domain names should be rendered using only U+002E as the label separator.

A similar argument can be developed for the display, storage and “on the wire” exchange of domain names made up exclusively of canonical labels. That is, labels using only PVALID characters (this is a term of IDNA2008 art).

Another principle expressed in IDNA2008 is that A-labels (these are ACE coded objects prefixed with “xn- -“) and U-labels (made up exclusively of PVALID characters taken from the Unicode suite of characters) are complementary. That is, each label can be translated into the other form by the Punycode algorithm plus or minus the pre-fix string required by the A-Label form. They are 1:1 in relation to each other.

An advantage of sticking to such a canonical practice is that the display and exchange of domain names becomes completely consistent regardless of the use of the U-label or A-label form.

One problem with widespread mapping is that the unmapped objects are sometimes treated as if they are valid (canonical?) forms, even when they are not. We have seen the problem this has raised with the sharp-S character. At the time IDNA2003 was adopted, there was no upper case sharp-S. The practice in IDNA2003 was to use Unicode CaseFold to map lowercase sharp-S into upper case “SS” and then map that into lower case “ss”. The consequence of widespread uniform mapping was to introduce lowercase sharp-S into the lexicon of “valid” labels even though no such label could ever be registered. Under IDNA2003 practice, a label with the “ss” substituted for “sharp-S” was registered. Users were able to enter the label with the lowercase sharp-S successfully because of mapping.

Then the Unicode Consortium added uppercase sharp-S into the Unicode tables. If one followed the principle of least astonishment, this should have resulted in lowercase sharp-S casefolding into lowercase sharp-S and uppercase sharp-S casefolding into lowercase sharp-S. However, because of the earlier mapping behavior, previously registered “ss” forms would not have been found. So by exception, sharp-S was casefolded into “ss” rather than lowercase sharp-S.

The IDNABIS working group has gone around and around about how to deal with this problem. Introduction of sharp-S as PVALID creates the conflict between the earlier mapping practice and the use of only canonical forms. A principle of the canonical form notion is that no PVALID character should be mapped into another PVALID character for obvious reasons (it introduces ambiguity). There is no ambiguity if lookup and registration are limited to canonical, PVALID representations only.

Another problem with assuming that uniform mapping behaviors can be adopted for all uses of domain names is that domain names show up in a wide range of contexts, not all of them web-related. These various contexts can fall afoul of assumptions about standard mappings.

One of the objectives of the IDNABIS working was to produce rules that allow independence of the standard from particular versions of Unicode. This it has done reasonably well. The debates now center on the collision of the canonical IDNA2008 posture and the mapping behavior of IDNA2003. In an attempt to reconcile various working group views on this problem, a document [draft-ietf-idnabis-mappings-05.txt] was produced that, by consensus, was not considered normative. It suggested that certain kinds of mapping on lookup (but not on registration) might be reasonable.

Taking into account that the IETF has rarely adopted positions on user interface design, it has been proposed that the resolution of inconsistencies between IDNA2003 and IDNA2008 with regard to mapping and related issue be moved outside the IETF and into a group with closer connections to UI design and broader representation from the relevant communities, including representation from the DNS operational community and those who actually have to manage registrant and user expectations during the transition and that this be done after IDNA2008 is confirmed and constitutes a stable basis for discussion.

An argument against this proposal seems to be that the Unicode Technical Committee Technical Report TR46, which advocates certain kinds of mapping behaviors, has not been reconciled with the canonical forms recommended by IDNA2008.

At least one informal proposal (not even an I-D; see appendix A) has been put forward that would allow registries (at all levels of the DNS) to independently and asynchronously introduce the canonical use of “problem” characters (ie formerly mapped by IDNA2003 into other now PVALID characters). This particular proposal treats as “harmless” the case where the adoption of canonical forms-only behavior may result in an NXDOMAIN response while previous mapping behavior would produce a valid DNS record. Several other proposals have surfaced involving “bundling” (concurrent registration of multiple “equivalent” forms). There is no apparent consensus within the working group as to which tactics are preferable and I would speculate that there may be different answers depending on the context of use of the domain names produced under IDNA2008 rules. Moreover, there is much evidence that further extended discussion within this working group will not converge.

A strong argument can be made that the job of the IDNABIS working group is to produce a standard way of using the Unicode character set that is consistent and version independent. I believe it has done this. The question of applying this canonical form in practice might reasonably be taken up by the groups with domain expertise (no pun intended). This includes wrestling with the question of the potential incompatibility of the IDNA2003 and IDNA2008 practices.

Working Group Question:

1. Would the WG like to adopt the current “mapping document” as-is
2. Would the WG like to engage in further discussion about this document, for example in the context of the Unicode TR46 that advocates substantially more mapping than the present “mappings” document?
3. Would the WG propose an alternative path towards dealing with the question of mapping and if so, what proposition(s) are offered by the WG members?

APPENDIX A

Introduction of Eszett (sharp-S) and Final Sigma

See http://typefoundry.blogspot.com/2008/01/esszett-or.html for an interesting perspective on 'Sharp-S'

Introduction

The IDNABIS working group has spent two years evolving documents describing the use of Unicode in Internet domain name labels. We have ended the IETF Last Call with a lengthy discussion on the manner in which the Unicode characters Latin Small Letter Sharp-S (U+00DF) and Greek Small Letter Final Sigma (U+03C2) are to be introduced into use. The so-called Zero-Width Joiner and Zero-Width Non-Joiner (ZWJ and ZWNJ respectively) have been included as CONTEXT-Joiner (or CONTEXTJ) in the IDNA2008 documentation and the general consensus is that these two may be registered at the discretion of registries. IDNA2008 specifically permits their use, in context.

The primary debates surrounding Sharp-S and Final Sigma relate to the method of their introduction into use as PVALID characters under IDNA2008. This note represents an attempt to synthesize a philosophical basis for achieving the goal of making these two characters usable in domain name labels.

It is useful to recall that the Domain Name System is a hierarchical system of registries. The root zone is the place where top level domain labels are registered. The Top Level domain name registries (e.g. .com, .coop, .ca, .uk) are 'pointed to' using 'delegation records' in the root zone file. Each 'dot' in a domain name is a point where 'delegation' (in DNS-speak, a zone cut) for further registration handling MAY be implemented.

So, for example, suppose that it is desired to create a Second Level label, 'foo' under the Top Level Label 'com'. Typically, the party wishing to register domain names with the suffix 'foo.com' would request to register 'foo' as a second level label under 'com' and a delegation record would be created pointing to the name server that will respond to all domain names with the suffix 'foo.com'.

At any point, a registration may either be an address record for, e.g., abc.foo.com, or a set of delegation records pointing to the servers Third Level label 'abc'.

The notion of delegation is important to keep in mind when considering how to introduce new PVALID characters into labels since each label in a multi-label domain name can be managed by a different entity (ie through delegated authority). A decision by a higher level authority to treat two different labels as equivalent is a non-trivial exercise in delegation mechanics. This fact is often lost in discussions about domain names as if there were flat identifiers. They are not. They really represent delegated hierarchies and their creation is often achieved through a series of assignments of delegated authority.

DESIDERATA ON THE INTRODUCTION OF NEW PVALID CHARACTERS

1. It is desirable that they can be introduced as soon as any registry in the hierarchy wishes to do so without having to coordinate with other registries.

2. It is desirable that IDNA2003 compliant and IDNA2008 compliant entities (programs, applications, etc.) co-exist without introducing ambiguous resolution of domain names (ie. The same domain name resolves to different IP addresses under IDNA2003 and IDNA2008 interpretation)

3. In the proposal that follows, a relaxation of the constraint in (2) is that it is acceptable that IDNA2008 interpretation leads to NXDOMAIN even if IDNA2003 leads to a valid IP address (or vice-versa). Under this provision, the introduction of a new PVALID character does not lead to distinct IP addresses (and therefore hazardous ambiguity) even if it produces (temporary?) non-resolution for some cases.

It should be recognized that the millions of registries/zones in the DNS are largely independent entities. We can produce a "suggested good practice", but registries will make local determinations as to what to do based on local considerations. To discourage a particular practice, it seems best to explain what bad consequences will result from following it but as a practical matter leave the decisions up to the registry. In many ways we have already adopted this position in IDNA2008 by leaving a great many decisions about which characters to permit for registration (even if they are PVALID in protocol) for reasons of local significance or practice.

There are many side-effects associated with introducing as PVALID characters that were formerly mapped under IDNA2003. An unknown number of URLs (or other domain-name-referencing constructs) may become unreachable upon adoption of IDNA2008, if the unmapped versions of the associated domain names have not been constructively registered and made to resolve to the same IP address as the mapped version.

THE SHARP-S EXAMPLE

Under IDNA2003, any reference to a domain name label containing Sharp-S is converted to a label containing 'ss' in place of Sharp-S, whereever Sharp-S appears. This revised label is then used either for registration or look up in the Domain Name System.

Under IDNA2008, Sharp-S is treated as PVALID and not converted to 'ss'.

Many of the suggested transition tactics have attempted a kind of "perfection" in which there is either a deadline by which everything works under IDNA2008 or new mechanisms to somehow distinguish between IDNA2003 and IDNA2008 or urge strenuous efforts to make everything backward compatible with IDNA2003 mappings - especially for the two problem characters Sharp-S and Final Sigma. I am ignoring everything else but these in this contribution since my sense is that this working group may go along with anything that "solves" the problem with them. Joiners I think we can assume have been accepted in the CONTEXTJ form.

I would like to try out on you an idea that isn't "perfect" but that avoids the worst hazard, I think.

My definition of worst hazard is that different entities (browsers, applications) do resolution and get conflicting results.

An example of this would be a case where under IDNA2003, a domain name containing Sharp-S would be vectored to a domain name and associated IP address that referenced a domain name registered with "ss" in lieu of Sharp-S and under IDNA2008 would be vectored to an IP address associated with a Sharp-S registration that leads to a different IP address and a distinct registrant. I would distinguish this from the case where the same registered domain name is associated with two or more IP addresses on purpose (e.g. two A records that the registrant considers equivalent).

IDNA2003 Case

registered	looked up		
domain name	domain name	IP address	Registrant
masse.com	maße.com mapped to masse.com	12.34.56.78	Mr. Foo

IDNA2008 Case

registered	looked up		
domain name	domain name	IP address	Registrant
maße.com	maße.com	34.56.78.12	Mr. Bar

The hazard is that under IDNA2003, a look up for maße.com gets the 12.34.56.78 address of masse.com while under IDNA2008, the look up for maße.com gets the 34.56.78.12 address of maße.com

What we would like is to prevent this unexpected ambiguity.

I would like to introduce a failsafe practice that prevents this particular ambiguity but allows for an NXDOMAIN result that may not be considered hazardous even it is annoying.

Let us imagine that the .com registry wishes to introduce IDNA2008 capability into its second level domain registrations (that's all it controls).

We assume that it has been registering under IDNA2003 rules in the past, so that any label containing "ß" will have been mapped to "ss" prior to registration. There is a collection of registrants in the equivalence class "registered a label containing 'ss'". Let us call the set of such registrants R.

The .com registry introduces a sunrise period in which all members of R are advised that they may register domains equivalent to the ones they did register but with the mapped "ss" form changed to the unmapped "ß" form. I am pretty sure there cannot be collisions here because all the final registrations have to have been mapped to "ss" - so if there were going to be a collision it would already have been detected at the time of original IDNA2003-compliant registration: "sorry, someone else has already registered the 'ss' form you would have gotten, can't register that."

After time T (determined by the registry, not by IETF or ICANN fiat), the .com registry then advises that it will accept registration of SLDs containing "ß". However, it abides by the following rules at REGISTRATION time:

(Failsafe Rule 1): If registration of an SLD containing "ß" would collide under IDNA2003 mapping rules with an existing registered domain name, the registration is allowed if the holder of the requested domain is the same (*) as the holder of the already-registered domain, otherwise the registration is not allowed.

(Failsafe Rule 2): If registration of an SLD containing "ss" would collide under IDNA2003 mapping rules with an existing registered domain name containing "ß" it is allowed if the holder of the requested domain is the same (*) as the holder of the already registered domain, otherwise the registration not allowed. Note that Failsafe rule 2 only applies once a registry is operating under IDNA2008 rules.

(*) Which registrants are "the same" is to be defined by the registry, and match the definitions the registry applies.

As a slightly less safe alternative, but at the option of the registry (perhaps after even more time has gone by), "not allowed" in the above two rules could be replaced by notification of the existing domain holder with an offer to again let that registrant preemptively register the name, thereby blocking its registration by someone else. If that offer were not accepted, the new registration would be permitted, of course still subject to whatever dispute resolution policies are in effect for .com or other relevant zone.

This latter suggestion opens the door for achieving independence of formerly-mapped pairs of now PVALID characters.

There are some nuances to the scenarios offered above. With possible exceptions for some "bundling" practices, most registrations will be sequential (ie. not "at the same time"). One typically registers one domain name and then registers others. Because of this, we will usually end up in a situation where at the time of the second (or Nth) registration someone has to check, for example, whether the requested holder of the next domain name registered is the same holder as the holder of earlier but colliding registered domain names.

There may be different registrars involved in sequential registrations. There may be different contact representatives for respective registrations. There might be transfers being made in between related registrations.

Because of this, the important things are the failsafe rules, and that they (in an ICANN context) are formulated by the registries so that details like "same" actually have some specific meaning in the specific registry context.

If we go back to the example given above and assume that Mr. Foo has registered masse.com before Mr. Bar has entered the picture, Mr. Foo will get to register maße.com during the sunrise period. Mr. Bar will not be allowed to register either maße.com or masse.com because both of these collide with previously registered domain names.

Let us now suppose that after the sunrise period, the registry is operating under IDNA2008 rules. Let us suppose that someone, Mr. Baz, has registered "strasse.com" prior to the adoption of the IDNA2008 rules. Let us also assume that he did not bother to register "straße.com" during the sunrise period (if he had, he would presumably have that registration too).

Now let Mr. Frotz try to register "straße.com" - under Failsafe Rule1, he would be denied this registration. Mr. Baz still has the possibility of registering it.

If someone looks up "straße.com" under IDNA2003-compliant rules, hewill get "strasse.com" unambiguously.

If someone looks up "straße.com" under IDNA2008-compliant rules, he will get NXDOMAIN. This is a kind of brokenness but perhaps this is tolerable if it does not steer the party to the "wrong" site - and it potentially allows Mr. Baz to recover from his earlier choice not to register the "ß" version of his SLD earlier.

Now let us suppose that "strasse.com" has NOT been registered at all, the sunrise happens, and we are now operating under IDNA2008 rules.

Mr. Frotz registers "straße.com". Since there is no collision with a previously registered "strasse.com" there is no problem. Let us suppose that Mr. Frotz does not bother to register "strasse.com".

If someone looks up "straße.com" under IDNA2003-compliant rules, he will get NXDOMAIN because "strasse.com" does not exist.

If someone looks up "straße.com" under IDNA2008-compliant rules, he will get the corresponding IP address.

If someone looks up "strasse.com" under IDNA2008-compliant rules, he will get NXDOMAIN because it has not been registered.

Because the registry is operating under IDNA2008-rules, "ß" and "ss" are considered distinct and the party using IDNA2003-rules to look up a domain name registered under IDNA2008 rules is getting a "correct" response in some sense (in this case, NXDOMAIN). At least the lookup does not lead to the "wrong IP address".

If Mr. Frotz registers both "strasse.com" and "straße.com" (assuming neither of these violates Failsafe Rules (1) and (2) at registration time), his registrations will work for both IDNA2003-compliant and IDNA2008-compliant lookups. Whether queries using the two strings will produce the same results or not will still be up to him and not the registry: there is no practical way to avoid that.

Let us suppose, again, that Mr. Frotz successfully registers "straße.com" under IDNA2008 rules but does not bother to register "strasse.com"

Now let us suppose that Mr. FUBAR tries to register "strasse.com" subsequent to Mr. Frotz's registration of "straße.com". When he tries to do this, he would be blocked from that registration under Failsafe Rule (2). Or, under the more permissive variation, Mr. Frotz would have an additional opportunity to block Mr. FUBAR's registration by registering "strasse.com" himself.

I believe that adoption of Failsafe Rules (1) and (2) would permit each registry (in the general sense - all levels) to introduce
IDNA2008 rules whenever they wish, and to provide for sunrise time periods of their choosing. The failures that occur (NXDOMAIN) are not harmful in the same way that "wrong IP address" would be harmful and perhaps this form of "failure" would be an acceptable price to pay for some period of time when IDNA2003-compliant and IDNA2008-compliant systems were in concurrent operation.

I hope this isn't completely nuts.

vint

from John Klensin:

The suggested process could be used to create a five-stage process:

(1) No registrations that actually involve Sharp-S (the status quo)

(2) Sunrise -- priority registrations for Sharp-S those who already have labels containing "ss".

(3) No possibly-conflicting registrations, using Failsafe Rules 1 and 2 as written; starting time to be determined by registry

(4) Possibly-conflicting registrations permitted only after the original registrant gets notification and an additional opportunity to register the name herself; starting date again determined by the registry

(5) Sharp-S is just another character with no special treatment; starting date again determined by the registry.