

IDN Working Group

Final report

[Draft 2006-02-16]

Introduction

The IDN Working Group was set up to consider how Nominet should handle IDNA – Internationalized Domain Names. After internal consideration a public consultation was held. This report gives the working group’s conclusions and proposals.

Terminology

Throughout this report “domain name” is generally used to refer to a domain name label (usually at the third level) in Nominet’s registry, rather than the full domain name (thus *sample* rather than *sample.co.uk*). Terms defined by Unicode or the IDNA RFCs may be used without being explicitly labelled as such.

Summary

- Nominet should introduce IDNA as soon as practical.
- Reasonable support should be provided, but not “at any costs”.
- Domain names should be limited (with some technical exceptions) to a single script until experience has been gained.
- It should not be possible to register names differing only by accents, but a “domain alias” facility should be added to allow a registration to cover names differing only in accents.
- No sunrise period is necessary, and DRS should be able to cope.

The Consultation

The IDN consultation was open from 6 June 2005 to 6 September 2005. It asked 13 questions about the various aspects of IDNA and how Nominet should handle it. There were a total of 46 responses and, in addition, some discussion on a PCPro bulletin board. The full consultation and responses are available from the Nominet Executive and will not be repeated here.

Should IDN happen and how?

About a third of respondents, and most of those taking part in the PCPro discussion, felt that Nominet should not allow IDN at all. The two main reasons given were:

- it would provide opportunities for abuse through “typo-squatting” (e.g. by using Greek letters instead of Latin ones to mimic existing names);
- it would require registrants to register numerous variations of their domain name for no benefit in order to prevent such typo-squatting (e.g. Nestlé would have to register all of *nestle.co.uk*, *nestlé.co.uk*, *nestlè.co.uk*, *néstle.co.uk*, and so on).

The remaining respondents felt that Nominet should move into this area. The reasons given generally agreed with those given in the consultation: it allows registrants to use the “natural” form of names rather than having to distort or transliterate them into the LDH character set.

Opinion varied on whether Nominet had some kind of obligation to provide IDN (e.g. under the Welsh Language Act). Inter alia, this could depend on whether it has “public body” status in this context. However, whether or not it does (and Nominet’s own legal advice is that it doesn’t) IDN would be in the interest of at least some stakeholders and should be considered on its merits.

For those who were in favour of introducing IDN, the strongest support was for option 3 (automated support facilities but no requirement for major process changes or new facilities). It was noted that additional facilities could be introduced as demand requires and money permits. Option 3 may in fact be cheaper than option 2, even though it involves more up-front effort, because it will reduce the scope for confusion and support problems.

Conclusions

The Working Group believes that Nominet should offer IDN to registrants, subject to the other conclusions in this report, as soon as reasonably practical.

As a general policy principle, IDN should affect tag-holders and registrants who do not take advantage of it as little as possible. It is appreciated that, particularly for tag-holders, the impact will not be nil. Nevertheless, the benefits to those registrants using IDN will, in the Working Group’s opinion, outweigh the disadvantages.

The implementation should, as far as possible, be provided through automated systems and education and information for both tag-holders and registrants, rather than by increased operational costs which would ultimately flow to all domain name holders. Nominet should, for instance, look to provide conversion tools for public use and toolkits for tag-holders to help them implement IDN. Education should include warning registrants that not all third-party systems and software may properly support IDN. Nominet support staff would not be expected to have a knowledge of a large range of character sets and would only be required to deal with ACE labels. This would, of course, not prevent tag-holders from providing further support to their customers, and the Working Group noted that this could be a viable niche market.

Nominet should adopt a pragmatic approach to cost-control. IDN should not be treated as a separate accounting centre that has to recover its own costs, but equally – as with any other part of the business – facilities should not simply be provided regardless of cost. Nominet should also ensure that there is no disproportionate cost or burden to existing stakeholders.

Avoiding typo-squatting

As mentioned above, a recurring theme in responses was the issue of typo-squatting. For example, the domain name *voμivet.org.uk* may appear innocuous as written, but when capitalised it becomes NOMINET.ORG.UK, but written with Greek rather than Latin characters. This would be a separate name to *nominet.org.uk*. More dangerously, both *nominet.org.uk* (with the first “o” actually being an omicron) and *nomīnet.org.uk* (note the two dots over the “i”) are separate names as well. There are literally millions of ways to modify names in similar ways, and no registrant should be expected to make lots of defensive registrations at no benefit to themselves.

While not all typo-squatting can be avoided (consider *nom1net.org.uk* or *NOMINET.ORG.UK*, the latter using digit 0 instead of letter O), the Working Group considered that it was

reasonable to put restrictions on names – over and above those in the relevant RFCs – to prevent the most common abuses. They recommend two restrictions over and above those in the RFCs (that is, to be allowed to be registered, names must conform to the RFCs and both these additional restrictions).

- Firstly, domain names are intended to be just that – names – and therefore they should be limited to the characters normally found in names, including digits and interior hyphens. Other punctuation marks, currency symbols, spaces, and “dingbats” such as ♠ should be forbidden; it is also noted that they do not form part of the “natural language” of any minority group.
- Secondly, domain names should be limited to a single script (e.g. Latin, Cyrillic, or Arabic) rather than being allowed to mix scripts. Mixed scripts (with a few specific exceptions) are more likely to be confusing or used for abuse. [This rule is based on policy option 2 from Unicode Technical Report 36 and has been adopted by other registries.]

The Working Group felt it was better and easier to remove rules that turn out to be unnecessary than to attempt to add new restrictions at a later date. Therefore, once IDN has been active for a reasonable time (say 6 months) and practical experience has been gained, Nominet should review the application of these restrictions and consider whether it is reasonable to relax them.

Some respondents proposed that domain names should be limited to certain scripts, such as the ones used for languages commonly written in the UK. The Working Group did not consider any of the reasons given to be convincing. However, they did believe that Braille was – unlike most scripts in Unicode – a variant of ASCII and not a separate script. As it turns out, Unicode considers the Braille script to be symbols rather than letters.

In addition, even if the “aliases” proposal described below is not adopted, the rule marked ‡ concerning unmarked names should be applied.

Technical proposal

The Working Group proposes that domain names should meet the following requirement in addition to those in the relevant RFCs. These rules apply after the name has been passed through the Nameprep algorithm.

- All characters must belong to one of the Unicode General Categories:
 - Letter (this includes Chinese ideograms and similar characters)
 - Number
 - Mark (these are mostly modifiers like ^, not symbols like £)or (except for the first or last character) be U+002D (“hyphen-minus”).
- Ignoring any characters with the Script property “Common” or “Inherited”, the remaining characters must either all have the same Script property, or must all have Script properties listed in the same one of the following sets:
 - Bopomofo, Han
 - Han, Hiragana, Katakana
 - Han, Hangul(Unicode currently has 59 possible Script properties – listed in Annex A – plus the special cases “Common” and “Inherited”.)

Modifier marks

An issue that has been raised frequently during the IDN consultation is how to handle “modifier marks”, also known as “diacriticals”, “accents”, “combining marks”, and other terms. These are the small marks that are placed above, below, or beside major characters to modify their meaning. The best-known, to Western European eyes, of these are the accents found in languages like French, German, and Spanish, such as the ´ and ~ found in é, ü, and ñ. However, Unicode has many more (863 in total).

Modifier marks lead to three problems:

- It is easy to confuse names that differ only in the modifier marks used (e.g. the double acute ¨ looks very similar to the diaeresis ¨, particularly in smaller fonts).
- Organisations with accents in their names may have already registered the unaccented version.
- Where an accented version is registered, users may enter the unaccented form through confusion or lack of support for accents in their tools.

On consideration, the Working Group concluded that different registrants should not be able to register names differing only in the modifier marks used. One way to do this would be to introduce some concept of “linked registrations”, but the Working Group felt that this would introduce too many new issues and violate the principle of minimal impact described above (for example, what if the names were registered on different dates or managed by different tagholders). Instead, the Working Group created the idea of “domain aliases”. Under this proposal, registrants would be able to create a small number of aliases (say 5) of a registered domain name, differing only in the modifier marks used, and all would operate identically: a DNS request for an alias would be answered as if it were for the main name (one way to do this might be the DNAME facility). Alias management would be done by the tag-holder or end registrant without Nominet’s intervention and Nominet would not charge for it. This would remove most, if not all, of the need for defensive registrations in relation to modifier marks.

Technical proposal

- The “unmarked form” of a domain name is found by applying Nameprep, then normalising using NFD, then removing all characters with General Category “Mark”.
- No two registrations in the same SLD shall have the same unmarked form. ‡
- All domain names (including those using only LDH characters) shall be permitted a fixed, limited, number of “domain aliases”. The domain aliases and the domain name shall have the same unmarked form.
- As far as technically practical, domain aliases shall behave the same as the domain name; in any case, the technical behaviour of aliases shall be clearly documented.
- Tools such as “whois” and the Domain Availability Checker should correctly handle the cases where a query is for:
 - a domain alias of a registered name;
 - a name which is neither registered nor an alias, but has the same unmarked form as a registered name.
- Nominet should consider providing a facility to change the domain name to one of the existing aliases, but this is not an essential part of the proposal.

[Note: the Working Group chose not to require the domain name to be the “unmarked form” because it may well have a significantly different meaning to the name the registrant actually

wishes to use. In particular, they were made aware of at least two instances where the unmarked form of a word had sexual connotations while the accented form did not.]

Sunrise Period

Given that the proposal for aliases will allow many registrants to reserve their desired name right now simply by registering the unmarked form, the Working Group does not believe that a sunrise period will be necessary. This will not assist those whose desired name does not have an unmarked form using only LDH characters, but the group does not think there will be a significant rush of registrations on day 1.

However, the Nominet Executive may wish to consult the operational and technical departments on this matter before reaching a final conclusion. If they decide that a “sunrise” period is necessary, it should involve a premium price over the first few months, gradually reducing to the normal price (just as happened with *me.uk*).

Dispute Resolution

The Working Group did not see the need for any special treatment of IDNs in the DRS process. Where a claim turns on some property of a name that would not be visible to the typical English-speaking person, the parties would be able to adduce expert evidence on the topic at the time of submission of the claim or defence. This would allow the DRS expert to make a decision even though she is not fully acquainted with the property in question, just as a judge decides between conflicting expert testimony in court. To this extent, cases turning on IDNs would be no different than those addressing any other specialized vocabulary.

However, the Working Group recommend that Nominet consult its DRS experts on this matter, as they are likely to understand the issues better.

Other matters

One respondent pointed out that Welsh companies should be using SLDs *cyf.uk* and *ccc.uk* instead of or as well as *ltd.uk* and *plc.uk*. While this is outwith the scope of the consultation, the Working Group notes that the DNAME method suggested for domain aliases could also be used here.

Finally, the Working Group encourages Nominet to take a full part in dealing with the international issues of IDN in the future, including working with other registries, the IETF, and the browser development community. While other countries may have greater experience with other character sets, the UK is possibly unique in having a significant minority market for IDN within a largely ASCII-based culture.

Annex A – Unicode Script types

Unicode recognises the following 59 script types, plus two special categories of “Common” and “Inherited”. Future types may be added in revisions to the Unicode standard. Those types in italics are treated specially in this report.

Arabic	Gurmukhi	Old Persian
Armenian	<i>Han</i>	Oriya
Bengali	<i>Hangul</i>	Osmanya
<i>Bopomofo</i>	Hanunoo	Runic
<i>Braille</i>	Hebrew	Shavian
Buginese	<i>Hiragana</i>	Sinhala
Buhid	Kannada	Syloti Nagri
Canadian Aboriginal	<i>Katakana</i>	Syriac
Cherokee	Kharoshthi	Tagalog
Coptic	Khmer	Tagbanwa
Cypriot	Lao	Tai Le
Cyrillic	Latin	Tamil
Deseret	Limbu	Telugu
Devanagari	Linear B	Thaana
Ethiopic	Malayalam	Thai
Georgian	Mongolian	Tibetan
Glagolitic	Myanmar	Tifinagh
Gothic	New Tai Lue	Ugaritic
Greek	Ogham	Yi
Gujarati	Old Italic	