ICANN | RrSG Registrar Stakeholder Group

RySG gTLD Registries Stakeholder Group

ICANN | IPC Intellectual Property Constituency

# Internationalized Domain Names, Unicode Homoglyphs and Abuse Mitigation Approaches

## 1. Overview

Internationalized Domain Names (IDNs)[1] enable Internet users around the world to use domain names in their own script and language of choice, enhancing linguistic diversity across the Internet and expanding opportunities to identify themselves in unique and memorable ways. This is made possible in great part through the efforts of the Unicode Consortium[2] to document and develop a standard encoding system for virtually all script characters from virtually all languages, which continues to evolve to include additional characters with each new version, and a forward-looking set of IETF volunteers who ensured that these new types of domain names would be compatible with core Domain Name System (DNS) standards. While every Unicode character has a unique code point that makes it immediately distinguishable from all other characters in automated systems, some Unicode characters from one script can be, in certain circumstances, visually similar to one in another script which can lead to user confusion[3]. When this type of visual similarity is abused, such a threat is known as script spoofing or IDN homoglyph attack. Over the years, guidelines and practices have been developed to balance the availability and utility of IDNs on one hand, and security concerns on the other. This document describes a collection of approaches and guidelines and the motivation for their use in domain name registration systems.

## 2. Purpose of this Document

The purpose of this document is to provide the reader a high-level understanding of the problem space and practices —some used over many years— which have been implemented to balance the availability and utility of IDNs on one hand, and security concerns on the other. This document is jointly drafted by the Registry and Registrar Stakeholder Groups and the Intellectual Property Constituency, which aim to promote the adoption and usage of IDNs that balances product usefulness and security and stability.

[1] IDNA2008 RFCs series (RFC5890, RFC5891, RFC5892, RFC5893, RFC5894, RFC5895)
[2] https://home.unicode.org/
[3] https://www.m3aawg.org/sites/default/files/m3aawg-unicode-tutorial-2016-02.pdf

## 3. IDN Homoglyphs

IDN homoglyphs are two or more characters, or glyphs, that depending on the context in which they are presented —typeface, font style, font size— can appear identical or nearly identical. For example, Table 1 shows the visual similarity among letters in the Latin, Cyrillic and Greek character sets.

Table 1: Examples of IDN homoglyphs[4]

| Script | Latin | Cyrillic | Greek |
|---|---|---|---|
| Glyph | **a** | **а** | **α** |
| Description | Small Letter A | Small Letter A | Small Letter Alpha |
| Code Point | 0061 | 0430 | 03B1 |
| Glyph | **o** | **о** | **ο** |
| Description | Small Letter O | Small Letter O | Small Letter Omicron |
| Code Point | 006F | 043E | 03BF |
| Glyph | **j** | **ј** | **ϳ** |
| Description | Small Letter J | Small Letter Je | Small Letter Yot |
| Code Point | 006A | 0458 | 03F3 |

The typeface and/or font that is used to display a domain name is controlled at the application layer, e.g., web browser, website or email client —DNS is glyph-agnostic, because it is ASCII only. Therefore, as it is noted further in this document, applications play a role in making sure the end-user is presented with an accurate representation of a domain name, for Universal Acceptance and security reasons.

## 4. IDN Homoglyph Attack

The IDN homoglyph attack is a means an attacker uses to deceive potential victims to visit a spoofed website or act on an email message, where the domain name or email address is using a spoofed domain name —utilizing non-ASCII characters.

[4] Source: https://www.unicode.org/Public/security/latest/intentional.txt and https://www.unicode.org/Public/security/latest/confusablesSummary.txt

These types of attacks are one of the many techniques phishers use[5]. To address this issue, DNS actors from throughout the Internet stack, from registries and registrars to web browser companies, have implemented certain practices to address the issue.

# 5. Collection of Approaches and Guidelines

Over the years, approaches and guidelines have been developed —throughout the internet stack— to balance the availability and utility of IDNs on one hand, and security concerns on the other.

## a. ICANN

The ICANN IDN Implementation Guidelines[6] provide registries with best practices regarding the registration of IDN labels at the second level. Most notably, these guidelines prohibit mixing of different scripts in a single label (i.e., *no-script-mixing* rule) at the second level. In practical terms it prohibits the registration of second-level domain name labels that mix code points (e.g., letters) from different scripts —such as Cyrillic and Arabic. However, certain widely accepted combinations may be allowed, e.g., to support the Japanese language, which uses Katakana, Hiragana, Han, and Latin characters[7].

For example, let's consider the following labels:

example

| **Glyph** | e | x | a | m | p | l | e |
|---|---|---|---|---|---|---|---|
| **Code Point** | Latin E U+0065 | Latin X U+0078 | Cyrillic A U+0430 | Latin M U+006D | Latin P U+0070 | Latin L U+006C | Cyrillic E U+0435 |

This label is mixing Latin and Cyrillic code points in a single label, therefore it would not pass the *no-script-mixing* test.

[5] Global Phishing Survey: Trends and Domain Name Use in 2016, APWG, 26 June 2017, https://docs.apwg.org//reports/APWG_Global_Phishing_Report_2015-2016.pdf: "[p]hishers don't need to resort to such attacks. [...] the domain name itself usually does not matter to a phisher."
[6] The ICANN IDN Implementation Guidelines are incorporated by reference in gTLD Registry Agreements. As of the date of this document, version 3 is enforced. There is a proposed version 4 that is under review.
[7] See Restriction Level Detection, Highly Restrictive definition, Unicode UTS#39, http://www.unicode.org/reports/tr39/#Restriction_Level_Detection

пример

| Glyph | п | р | и | м | е | ρ |
|---|---|---|---|---|---|---|
| Code Point | Cyrillic PE U+043F | Latin P U+0070 | Cyrillic I U+0438 | Cyrillic EM U+043C | Cyrillic E U+0435 | Greek PHO U+03C1 |

This label is mixing three scripts, Cyrillic, Latin and Greek in a single label, therefore it would not pass the *no-script-mixing* test either.

While script commingling prohibitions prevent the registration of a large number of mix-script confusable IDN homoglyph domain names, the current IDN Implementation Guidelines do not address single script confusability. Consider the label:

exαmple (not example)

| Glyph | e | x | α | m | p | l | e |
|---|---|---|---|---|---|---|---|
| Code Point | Latin E U+0065 | Latin X U+0078 | Latin Alpha U+0251 | Latin M U+006D | Latin P U+0070 | Latin L U+006C | Latin E U+0065 |

This label does not commingle scripts and is therefore not prohibited by the IDN Implementation Guidelines. The “alpha” character is different from the letter “a”, yet also exists in the extended Latin script. To address this type of issue some registries adopt other voluntary measures, including those described below.

## b. IDN Variant Policies

The IDNA protocol and ICANN IDN Implementation Guidelines —collectively— define an operational framework that promotes a safe and stable implementation of IDNs. However, TLD registries implement additional voluntary policies that best meet their target market, including about *IDN variants*.

The concept of *IDN variants* was introduced to incorporate linguistic considerations to hostnames, such as the case of Chinese and its two writing systems —Simplified and Traditional. However, there is not a standard definition of *IDN variant*. One definition states an IDN variant “is an alternate code point (or sequence of code points) that could be substituted for a code point (or sequence of code points) in a candidate label to create a variant label that is considered the

'same' in some measure by a given community of Internet users"[8]. The relationship could be one based on visual, phonetic, or semantic grounds[9]. For example:

**Semantic Variants**: The Proposal for a Chinese Root Zone LGR[10] defines variants as "characters with different visual forms but with the same pronunciations and with the same meanings as the corresponding official forms in the given language contexts."

E.g., CJK Unified Ideograph, U+9093 (邓) is a variant of CJK Unified Ideograph, U+9127 (鄧)

**Phonetic Variants**: The Proposal for Ethiopic Script LGR for the Root Zone defines variants based on sound and meaning —i.e., code point redundancy or phonemic decay[11]

E.g., Ethiopic Syllable HA, U+1200 (ሀ) is a variant of Ethiopic Syllable HHA, U+1210 (ሐ)

**Visual Variants**: The Proposal for a Cyrillic Script Root Zone LGR[12] determined that there are no variant characters in Cyrillic script. However, it proposed variant relationships across other scripts, e.g., Latin, for those pairs with "identical graphemes" — i.e., homoglyphs.

E.g., Latin Letter A, U+0061 (**a**) is a variant of Cyrillic Letter a, U+0430 (**а**)

TLD registries use the IDN variant construct[13] as an additional method to minimize the registration of homoglyph labels. Since there is no single standard for IDN variant code points, registry implementation can vary[14]. Each registry can define its variant rules, or choose one of many documented practices, depending upon its business model, target market, legal framework, etc.[15]

[8] Definition of IDN Variant, LGR Procedure 2013, https://www.icann.org/en/system/files/files/lgr-procedure-20mar13-en.pdf
[9] RFC8228, Variant Relations, https://www.rfc-editor.org/info/rfc8228
[10] Proposal for a Chinese Root Zone LGR, https://www.icann.org/en/system/files/files/proposal-chinese-lgr-26may20-en.pdf
[11] Proposal for Ethiopic Script LGR for the Root Zone, https://www.icann.org/en/system/files/files/proposal-ethiopic-lgr-17may17-en.pdf
[12] Proposal for a Cyrillic Script LGR for the Root Zone, https://www.icann.org/en/system/files/files/proposal-cyrillic-lgr-03apr18-en.pdf
[13] Eurid: Domain Names with Special Characters (IDN), accessed 26 Aug 2021, https://eurid.eu/en/register-a-eu-domain/domain-names-with-special-characters-idns/
[14] Variant relationship rules and disposition of variant labels is registry specific. This policy may include blocking variant labels from registration or reserving variant labels for the same registrant.
[15] gTLD registries are required to publish its IDN tables or Label Generation Rulesets (LGRTs) in the IANA Repository of IDN Practices, https://www.iana.org/domains/idn-tables

For example, ICANN developed a label generation ruleset (LGR) for the root zone[16]. This LGR contains a repository of IDN variant code points developed for a conservative zone such as the root.

## c. Registration Policies and Services

Some registries and registrars offer policies and services to protect domain names, especially names closely related to brands. For example, some registries offer services to reserve or otherwise block the registration of related names to prevent look-alike variations from registration by third parties. Another example is registries offering to package multiple related names to be registered along with a desired domain name. The related names often include similar strings, misspellings, domain names across TLDs, and IDN homoglyph names.

## d. Post-registration Processes

In general, when a domain name is identified in connection to a possible case of abuse there are steps a registrant or end user can take to address the issue:

IDN homoglyph attacks should be reported to the appropriate party, typically the web host or registrar. In certain circumstances, for example when the web host or registrar are not responsive, it may be appropriate to report the DNS Abuse issue to the registry. Registrars and Registries may have varying jurisdictional requirements to collect data to act upon a report, so it is important to confirm what the notice requirements are before filing a report.

In addition to reporting the IDN homoglyph attack to the registry or registrar, an action available when the attack leverages a trademark to conduct abuse, is to seek relief through the Uniform Domain-Name Dispute Resolution Policy (UDRP)[17]. Similarly, the Uniform Rapid Suspension (URS)[18] process may offer a lower-cost, faster path to possible relief for trademark owners experiencing clear-cut cases of infringement.

## e. Web Browser Practices

Web browsers, i.e., application layer, typically control the display behavior of domain names in their user interfaces, for instance, the domain name displayed in the URL address bar. With IDNs, web browsers process the transformation of the Unicode label to an ASCII compatible string for DNS resolution and do the reverse transformation for presentation purposes to the end-user.

[16] ICANN's Root Zone Label Generation Ruleset, https://www.icann.org/resources/pages/root-zone-lgr-2015-06-21-en
[17] ICANN UDRP, https://www.icann.org/resources/pages/help/dndr/udrp-en
[18] ICANN URS, https://www.icann.org/resources/pages/urs-2014-01-09-en

Over the years, web browsers have implemented different strategies to strike a balance between utility and visual confusion. Some of these web browser's practices are:

i. Google's Chrome IDN Policy[19]
Chrome determines whether to show the user the U-label or A-label[20] for each domain label based on a set criteria, including flagging certain script combinations, checking for invisible characters and patterns, etc. For example, if a domain name is mixing Latin and Cyrillic letters, Chrome will show the A-label (e.g, "xn--exmpl-5ve2b" instead of "example" which is using Cyrillic "a" and "e" mixed with Latin letters).

ii. Mozilla[21]
Mozilla uses a combination of a whitelist for TLDs (their legacy solution) and an algorithm based on the Unicode Technical Report's "moderately restrictive" identifier profile — which, in general, prohibits commingling of Latin, Cyrillic and Greek letters in a single identifier.

# 6. Other Resources

## IANA

The Internet Assigned Numbers Authority (IANA) serves, among other functions, as a repository[22] of IDN registry policies. ICANN generic Top Level Domain registries are required to publish their IDN tables for second level domain name registration to provide transparency as to specific characters that are available within a script for registration, as well as contextual and variant rules, if any. These rules are composed by the repertoire of code points, variant rules (if any), and other evaluation rules to validate a string. Registries are encouraged to consult this repository to become aware of common practices and existing tables to help inform them as they determine the practices that best meet their target markets.

## M3AAWG

M3AAWG is the Messaging, Malware, Mobile Anti-Abuse Working Group, an industry association that addresses problem such as botnets, malware, spam, viruses, DoS attacks and other online exploitation[23]. This group published two papers to address issues regarding Unicode abuse: an

[19] Google's Chrome IDN policy, https://chromium.googlesource.com/chromium/src/+/main/docs/idn.md
[20] An "A-label" is the ASCII-Compatible Encoding (ACE) form of an IDNA-valid string. A "U-label" is an IDNA-valid string of Unicode characters, in Normalization Form C (NFC) and including at least one non-ASCII character, expressed in a standard Unicode Encoding Form (such as UTF-8), https://datatracker.ietf.org/doc/html/rfc5890#section-2.3.2.1
[21] Mozilla IDN Display Algorithm, https://wiki.mozilla.org/IDN_Display_Algorithm
[22] IANA's IDN Practice Repository, https://www.iana.org/domains/idn-tables
[23] About M3AAWG, https://www.m3aawg.org/about-m3aawg

overview and tutorial[24] and best practices for Unicode abuse prevention[25]. These best practices are targeted for a wide range of applications that process and display domain names and email addresses, and are not solely focused on the DNS.

## Unicode Consortium

The Unicode Consortium is responsible for developing the Unicode Standard. Unicode has been widely adopted as the character encoding system used at the edge of the internet, and other applications and protocols. The Unicode Technical Standard #39 discusses possible security problems —using Unicode in general and not just in domain names— and mechanisms to identify them, among these are:

- IDN Security Profile for Identifiers
- Detection Mechanisms to identify possible issues, such as problematic combinations of scripts in a single label

Also in Unicode, there is the Unicode Technical Report #36 which discusses the concepts of mixed-script spoofing and single-script spoofing, which some applications based their IDN policy on.

[24] Unicode Abuse Overview and Tutorial, M3AAWG, February 2016, https://www.m3aawg.org/sites/default/files/m3aawg-unicode-tutorial-2016-02.pdf
[25] Best Practices for Unicode Abuse Prevention, M3AAWG, February 2016, https://www.m3aawg.org/sites/default/files/m3aawg-unicode-best-practices-2016-02.pdf