Messaging, Malware and Mobile Anti-Abuse Working Group

M³AAWG Unicode Abuse Overview and Tutorial

February 2016

Executive Summary

This document examines the background of Unicode characters in the abuse context and provides a tutorial on the options that are emerging to curtail that abuse. Additionally, it discusses guidelines from the Unicode Consortium[1] that can be leveraged to standardize the abuse-fighting approach. Visually confusable Unicode characters – e.g., using the Greek letter omicron 'ο' in place of a Latin 'o' – have been used for many years to mislead users, but functional online elements like links and addresses were previously limited to ASCII so this type of abuse was limited.

However, with increased support for these characters in International Domain Names, Internationalized Top-Level Domains, and Email Address Internationalization, the extent and range of this abuse is poised to increase significantly. Network operators and related Internet-connected application operators can find Unicode anti-abuse guidelines in the M³AAWG Best Practices for Unicode Abuse Prevention document.

I. Background

International Domain Names (IDNs), Internationalized Top-Level Domains (TLDs), and Email Address Internationalization (EAI) allow for non-ASCII and non-Latin characters to be used in domain names and email addresses. Since 70 percent of Internet users speak languages utilizing non-ASCII characters, there is considerable underlying demand for URLs like "http://שלום.com" or "http://hellokitty.みんな" and email addresses like Jérome@example.fr. As users and systems increasingly support these non-Latin characters, the potential for abuse is rising.

The potential Unicode abuse comes not from these characters per se but from abusers taking advantage of the much-larger character set (there are more than 100,000 Unicode characters) to construct visually confusing sequences that mislead users and evade anti-abuse detection. For example, a user seeing a link to "https://Ьank.com" could easily overlook the fact that the first letter in "bank" is actually not a Latin 'b' but rather a Cyrillic capital letter "Soft Sign" ('Ь', Unicode: U+042C). Likewise, and even more subtle, is that in most computer typefaces, characters such as the Greek small letter "omicron" ('ο', Unicode: U+03BF) and Cyrillic capital letter "ve" ('В', Unicode: U+0412) are pixel-perfect duplicates to their Latin/ASCII equivalents 'o' and 'B'. (See Figure 1 below.)

Character	o	ο	о
Name	Latin small letter o	Greek small letter omicron	Cyrillic small letter o
Byte Sequence	0x006F	0x03BF	0x043E

Figure 1: In many fonts, these three characters appear identically.

Note that while Unicode "homoglyphs" – visually identical glyphs or symbols – are covered by this overview, this tutorial cannot completely cover all visually confusing combinations. Characters sequences that blur at small typefaces – like the Latin digit '1' and the Latin letter 'l', or digraphs like 'r' + 'n' blurring together to appear like an 'm', and regionally-directed Han characters like U+6B72 (歲) in Korean Han versus U+6B73 (歲) in Japanese Han – remain possible venues for deceivers. A future version of this document will address visual similarity in general.

II. Restriction Levels

For purposes of standardization, the Unicode Consortium has defined certain *combinations* of scripts as suspicious and unlikely to occur in natural language usage. For example, while a label may be written in any script or language, *switching* from Latin to Cyrillic *inside* the same label – as in the example "bank" above – is prohibited. These restriction levels are codified in Unicode TR39[2] as Highly Restrictive.

The Highly Restrictive definition specifies that all characters in each identifier must be from a single script or from certain specific combinations traditionally encountered in East Asian languages:

- *Latin* + *Han* + *Hiragana* + *Katakana*;
- *Latin* + *Han* + *Bopomofo*; or
- *Latin* + *Han* + *Hangul*

Under these restrictions, the following labels are allowed or disallowed:

- Allowed:
  - "José Üser" <joe@user.com> # All characters in Latin script
  - http://exámple.com # All characters in Latin script
  - http://みんな.example.com # 1st label all Katakana; 2nd and 3rd all Latin
  - http://example.みんな # 1st label all Latin, 2nd all Katakana
  - http://みんな 123.foo # Allowed combination of Latin + Katakana
  - ㄆㄎ-hello-両@foo.com # Allowed combo: Bopomofo + Latin + Han

- Disallowed:
  - http://www.goοgle.com # Greek omicron combined with Latin
  - "Joe User" joe@goοgle.com # Greek omicron combined with Latin
  - www.å∫†.ws # Mix of non-Latin scripts

Furthermore, the Highly Restrictive level specifies that characters in the identifier must all come from the "Identifier Profile," thus excluding emoji, characters not in modern use, characters only used in specialized fields (e.g., liturgical characters, phonetic letters, and mathematical letter-like symbols), and characters in limited use by very small communities. For example:

- Disallowed:
  - http://Τwitter.com # 'T' is actually archaic Greek letter Sampi
  - http://abc‌def.com # Zero-width space U+200B between chars
  - http://▫.com # Exotic character from 15th century poem[3]

As specified in Unicode Technical Standard #39, the IDN Security Profile for Identifiers[4] does not permit certain ASCII symbols and punctuation found in email addresses that do not appear in domain names, notably the ***dot-atom-text*** characters such as '+' and '&' from RFC 5322 §3.2.3[5] An effort is underway to add these to a new email address identifier profile in a future version of the standard[6].

III. Conclusion

The legitimate usages of Unicode characters are expected to grow rapidly with the advent of International Domain Names, Internationalized Top-Level Domains, and Email Address Internationalization. This document provides an overview of Unicode Consortium restrictive definition labels to help practitioners understand this abuse so they can define strategies and tactics to curtail its reach. Recommended best practices are detailed in the document, M[3]AAWG Best Practices for Unicode Abuse Prevention, also available from the M[3]AAWG website at www.m3aawg.org under Best Practices then select For the Industry.

IV. References

[1] The Unicode Consortium, http://unicode.org/

[2] Unicode® Technical Standard #39, 5.2 Restriction-Level Detection, http://www.unicode.org/reports/tr39/#Restriction_Level_Detection

[3] Multiocular O, https://en.wikipedia.org/wiki/Multiocular_O

[4] Unicode® Technical Standard #39, 3.2. IDN Security Profiles for Identifiers, http://www.unicode.org/reports/tr39/#IDN_Security_Profiles

[5] Internet Message Format RFC 5322, https://tools.ietf.org/html/rfc5322 - section-3.2.3

[6] Unicode Consortium, L2/15-080, Add email identifier profile to TR39, November 28, 2014 http://www.unicode.org/L2/L2015/15080-email-ident-profile.pdf

As with all best practices that we publish, please check the M[3]AAWG website (www.m3aawg.org) for updates to this document.