Messaging, Malware and Mobile Anti-Abuse Working Group

M³AAWG Best Practices for Unicode Abuse Prevention

February 2016

Executive Summary

For years, visually confusable Unicode characters – e.g., using the Greek omicron ('ο' U+03BF) in place of a Latin 'o' – have provided the potential to mislead users. Because most functional elements, like links and addresses, were previously limited to ASCII, abuse remained peripheral. However, the scope and definition of this abuse are poised to change with the advent of International Domain Names (IDNs), Internationalized Top-Level Domains (TLDs), and Email Address Internationalization as these non-ASCII and non-Latin characters gain in popularity.

This document outlines M³AAWG best practices to curtail the Unicode abuse potential of such spoofing, while supporting the legitimate use cases. The intended audiences for these practices are: email service providers, Internet service providers, and the operators of Software as a Service or others in relations to other Internet-connected applications.

A brief tutorial explaining how Unicode characters are used to perpetuate abuse can be found in M³AAWG Unicode Abuse Overview and Tutorial paper. This paper can be downloaded from the Best Practices section of the M³AAWG website.

I. Background

The M³AAWG best practices approach is to disallow certain characters not in common usage, as well as the combining of confusable scripts within a single label, with exceptions made only for certain visually distinct combinations found in legitimate usage. The intent of this exacting strategy is to inhibit suspicious combinations from taking hold by legitimate users, while ensuring a clean delineation of what constitutes abuse. For consistency and defensibility, these strategies are based on the Unicode Consortium's standardized "Restriction Levels" definition[1], which takes into account expected legitimate combinations of character sets.

M³AAWG recommends that service providers prohibit usage or advise users, or do both, of unacceptable combinations within three specific entity types: the **local part** of an email address, the **domain** of an email address, and the **URLs** found in clickable portions of a message or document. (See Figure 1 below.) Note that these practices do not cover the query-string portion of URLs, due to insufficient data on accidental and legitimate usage. As implementation of these practices increases, this document will be updated based on real-world data.

Although domain registries such as .com do place some limitations on allowed character combinations, significant abuse potential still exists, hence these best practices guidelines go further and limit a subset of those allowed by the registry policies. The names allowed by registry rules but disallowed here are unlikely to be of use to legitimate mailers or websites.

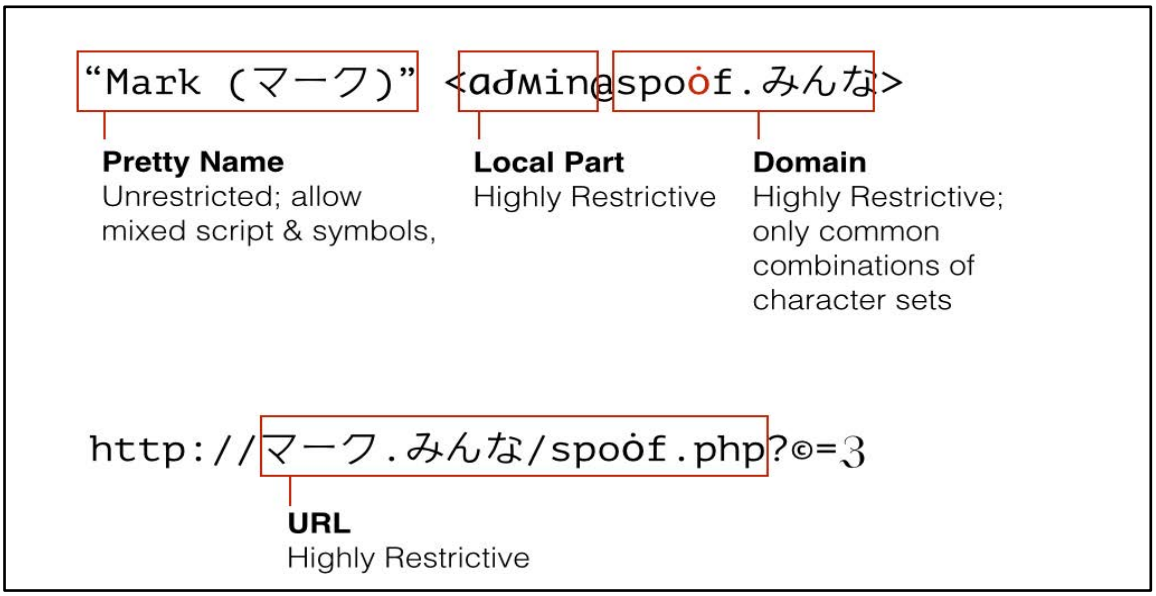

Figure 1: Components of email addresses and links and their best practices recommendations

II. Best Practices for Email

At **receive** time, MAIL FROM, FROM header, REPLY TO, and SENDER fields should be checked for validity and tested under the specifics in the Unicode TR39 Highly Restrictive level [1], as well as the normal RFC 5322 [2] and RFC 5321[3] syntax and validity checks. Messages with invalid addresses may be rejected or delivered to the spam folder.

At **send** time, in addition to the fields above, the TO, CC, BCC, and REPLY TO header recipients should also be checked under the same Highly Restrictive restriction level, with a send-time warning or block displayed on violation. See the Email Best Practices – Recommended Decisions table below.

Email Best Practices – Recommended Decisions			
	Condition	**Examples**	**Recommended Decision**
1	SMTP MAIL FROM, or FROM/SENDER header in DATA (the "FROM/SENDER Fields") contains a disallowed code point from the IDN Security Profile for Identifiers [4]	joe@foo{.com joe{@foo.com joe@foo{.com joe@foo.c{m	Inbound: Rejects with 5xx Outbound: Block

2	FROM/SENDER fields do not meet "Highly Restrictive" level [5]:
	• No "restricted" characters from "idmod" Identifier Profile[a]
	• All characters in each identifier must be from a single script or from the combinations:
	  • *Latin + Han + Hiragana + Katakana*;
	  • *Latin + Han + Bopomofo*; or
	  • *Latin + Han + Hangul*
		• joe@foo!.com *(U+01C3 "Latin letter retroflex click")*
		• joe@google.com *(Greek small letter omicron combined with ASCII)*
		• joe@google.com *(omicron again)*
		• joe@åʃ†.com *(mix of scripts)*
		• joe@foo.å@ʃ.† *(mix of scripts, invalid TLD)*
		• fooא.com *(mix of RTL and LTR scripts)*
			Inbound: Reject with 5xx
			Outbound:
			• If UI support: Block send with warning of suspicious copy
			• If no UI support: Reject message

3	FROM/SENDER (local part or domain) contains USPOOF_MIXED_NUMBERS[b]	joe@৪00.com *(U+09EA, Bengali digit four)*	Inbound: Reject with 5xx Outbound: Block
4	FROM/SENDER (local part or domain) contains a sequence of multiple non-spacing marks[c]	joe@ä̈t.com *(contains **both** U+00E4 Latin small letter 'a' with diaeresis **and a redundant** U+0308 combining diaeresis; in some fonts these are displayed overlapping)*	Inbound: Reject with 5xx Outbound: Block
5	BODY link	http://test.com *(contains Cyrillic small letter 'ie' (U+0435) along with Latin characters)*	Inbound: Disable link, display warning[d] Outbound: Warn with interstitial

Table Notes:

[a] http://www.unicode.org/Public/security/latest/xidmodifications.txt, and cf. the (less restrictive) Mozilla guidelines at http://kb.mozillazine.org/Network.IDN.blacklist_chars. Note that an effort is underway with the Unicode Consortium to create an identifier profile specific to email. If and when such proposal is ratified, implementers should update to follow it.

[b] See http://www.unicode.org/reports/tr39/#Mixed_Number_Detection.

[c] Non-spacing and combining marks are Unicode symbols that modify an adjacent character, such as an accent or a typographical hint to display a ligature. For more information, see http://unicode.org/faq/char_combmark.html and http://www.fileformat.info/info/unicode/category/Mn/list.htm

[d] As of press time for this document, insufficient data exist for a strong "block" recommendation on body links. Subsequent documents may endorse stronger policy.

III. Best Practices for Usage Outside of Email Messages

In addition to the usage in deceptive email addresses, numerous other opportunities exist within online services for similar exploitation, whether in URLs and links, display names, instant messaging addresses, account names or elsewhere. The breadth of use-cases is too much for this document to cover, but in general the recommendations follow those for users of the Email policy, i.e., that service providers should both disallow creating suspicious labels for active components and warn users before presenting these labels to them.

Best Practices Outside of Email – Recommended Decisions			
	Condition	**Examples**	**Recommended Decision**
1	Domain in URLs/links do not match "Highly Restrictive" level	`http://test.com` *(contains Cyrillic small letter 'ie' (U+0435) along with Latin characters)*	Disable link, with warning interstitial
2	Document names do not meet "Highly Restrictive" level [5]	• `Yahoo! Financial Information` *(U+01C3 "Latin letter retroflex click")* • `Login Instructions` *(Greek small letter 'omicro'n combined with ASCII)* • `Random åʃ†` *(mix of scripts)* • `fooא` *(mix of RTL and LTR scripts)*	• Disallow document name change • Warn user on attempts to open existing documents
3	Any labels use of USPOOF_MIXED_NUMBERS[e]	`8 00` *(U+09EA, Bengali digit four)*	• Disallow creation of suspicious label • Warn before displaying to users
4	Any labels using sequences of multiple non-spacing marks	`joe@ä t.com` (*contains **both** U+00E4 Latin small letter 'a' with diaeresis **and a redundant** U+0308 combining diaeresis; in some fonts these are displayed overlapping.)*	• Disallow creation of suspicious label • Warn before displaying to users

[e] See http://www.unicode.org/reports/tr39/#Mixed_Number_Detection

I. Conclusion

Deceptive homoglyphs have been used sporadically for years in messaging abuse, with limited success or adoption. As legitimate usages of Unicode characters rise with the advent of International Domain Names, Internationalized Top-Level Domains, and Email Address Internationalization, the potential for Unicode abuse increases accordingly. This document provides M[3]AAWG best practices to curtail this potential and ensure that abusive cases remain the fringe minority while enabling message receivers to take strong actions against illegitimate usage without fear of false positives. A general overview of Unicode abuse can be found in the paper M[3]AAWG Unicode Abuse Overview and Tutorial available at www.m3aawg.org by selecting For the Industry then Best Practices.

II. References

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

[2] RFC 5322, Internet Message Format, http://tools.ietf.org/html/rfc5322

[3] RFC 5321, Mail Transfer Protocol, http://tools.ietf.org/html/rfc5321

[4] Unicode® Technical Standard #39**,** Unicode Security Mechanisms, http://www.unicode.org/reports/tr39/

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

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

M3AAWG102