Image Hash, or, the last frontier of Emoji encoding

Steven R. Loomis (individual contribution)
srl@icu-project.org https://srl295.github.io

2016-04-29

1 Introduction

According to the Unicode Consortium's FAQ,

> Emoji are "picture characters" originally associated with cellular telephone usage in Japan, but now popular worldwide.[1]

The popularity of Emoji has been discussed elsewhere, but for the purposes of this document it is worth noting that over 735 code points have been designated Emoji[2] with more in the pipeline. The stated longer-term goal for Unicode, however, is that implementations would support *"embedded graphics, in addition to the emoji characters"*.[3]

This document aims to present a proposed implementation of said stated goal, by a mechanism for encoding a unique embedded graphic identifier in plain text. If such a mechanism were widely adopted, the need for additional code point allocation for Emoji should be obviated.

2 Non-Goals

UTR # 51 outlines some possible use-case scenarios as well as challenges with embedded graphics.[4] It is not the goal of this document to address all aspects of embedded

[1] The Unicode Consortium. *Frequently Asked Questions, Emoji and Dingbats*. 2016. URL: http://unicode.org/faq/emoji_dingbats.html.

[2] Mark Davis and Peter Edberg. *Unicode Technical Report 51: Unicode Emoji*. 2015. URL: http://www.unicode.org/reports/tr51/, Section 3, "Which Characters are Emoji".

[3] Ibid., Section 8, "Longer Term Solutions".

[4] Ibid., Section 8, "Longer Term Solutions".

graphics. This document will focus on those aspects related to character encoding only, and leave to domain experts and implementers to determine standardized approaches to topics such as privacy and security, actual data transfer of the image content, reliability and availability, and the like.

3 Overview

This document proposes:

1. The encoding of a new base character for image transfer, `U+FFF8 EMBEDDED IMAGE BASE`

2. The allocation of the entire plane `0C` for the purpose of image hashes

(TODO TODO)

To generate a hash, the image content (a standardized size, 128x128 png), plus the metadata (content-type, alternates, etc) is SHA-256 hashed.

The actual encoding is:

`U+FFF8 + U+0Cxxxx + U+0Cxxxx + U+0Cxxxx ...`

where each `U+0C` code point, from `U+0C0000 – U+0C7FFF` contains 15 bits of the SHA-256 hash.

The U+0C code points will have a combining character general category (which?).

The more `U+0C` present (up to 20 - 300 bits) the longer and more specific the hash is.

(TODO TODO)

References

Consortium, The Unicode. *Frequently Asked Questions, Emoji and Dingbats*. 2016. URL: http://unicode.org/faq/emoji_dingbats.html.

Davis, Mark and Peter Edberg. *Unicode Technical Report 51: Unicode Emoji*. 2015. URL: http://www.unicode.org/reports/tr51/.

Colophon

Typeset by LaTeX. Made with 100% recycled bits. All opinions belong to the authors and do not reflect the opinions of their associated employers.

Thank you to Keith Winstein for the discussion which finally kicked off this document.