L2/21-099

Re: **On Accumulated Feedback on QID**

From: **Mark Davis**

Date: **2021-04-28**

This document provides draft responses on the feedback to the QID proposal. The feedback was organized by general topics, with a response at the top of each topic section. The name and date for the feedback, sometimes with a short clipping, are included under the main topic of the feedback. The comments here do not include the full text of the feedback: see Accumulated Feedback on PRI #408 for full text. It may help to have that document open in a separate window while reading this one.

This document was produced after the last ESC meeting, and thus has not been reviewed by the ESC. The responses are draft, and may change after discussion in the UTC.

To get more context on QID, first read **Future Unicode Emoji Options (*L2/21-078*).**

Contents

Too Many Emoji

Plain Text Alternatives

Typos or wording fixes

Wikidata Reliability

General Support

Tangential

Identifiers, Duplicate Encoding, Confusability, Stability...

Image Transport

Too Many Emoji

"suddenly there are the potential to have thousands or millions of emoji based on these identifiers. This would create a significant burden for font developers"

This is not a change. Many people are not familiar with the fact that there are, ***right now***, an infinite number of valid emoji. Any sequence **emoji + ZWJ + emoji** (with **+ ZWJ + emoji** repeated as many times as one likes) is a valid emoji. And any company is free to make up new combinations of emoji ZWJ sequences.

The mechanism that the Unicode Consortium has put into place to handle this is that it distinguishes between the infinite number of valid emoji, and the relatively small number of sequences qualified as

RGI (Recommended for General Interchange). So “This would create a significant burden for font developers, especially as at launch none of these will be supported” is not a new issue, and one for which there is already a solution.

What QID does is just expand the constraints to allow for valid, customized emoji that can’t reasonably be represented by a ZWJ sequence. That allows for companies to use valid, but non-RGI, emoji in a way other companies can also pick up and use them interoperably . If the usage skyrockets, then that is a good signal to us that they could be candidates for RGI.

The difference from emoji zwj sequences is that there is a “rough” fallback for them. *NOTE: one commenter also noted an issue with having different TAG bases. The proposal no longer includes that capability; there would be only one TAG_BASE. That TAG_BASE effectively becomes the "missing emoji" glyph (also see below).*

> **Date/Time:** Fri Nov 22 15:19:23 CST 2019
> **Name:** Nicholas Felker
> **Report Type:** Feedback on an Encoding Proposal
> **Opt Subject:** Feedback on proposal #408 QID Emoji
> I want to provide some feedback on the proposal, as I think there are pros and cons to the approach. On one hand, I do like the effort to scale emoji to enable a broader set of pictorial characters. It would certainly enable novel and unpopular emoji to be used and shared.
>
> I think I share some of the concerns on others who have feedback. With this proposal, suddenly there are the potential to have thousands or millions of emoji based on these identifiers. This would create a significant burden for font developers, especially as at launch none of these will be supported. It would need to be incorporated into system-level keyboards and fonts. The OS vendors may be unlikely to support many of these in their font, which in turn would result in a lack of keyboard support and a lack in usage.
> ...
>
> **Date/Time:** Tue Nov 26 07:50:00 CST 2019
> **Name:** David Lewis
> **Report Type:** Public Review Issue
> **Opt Subject:** PRI #408: QID Emoji Sequences
> It seems to me that rather than supporting an entirely new mechanism for Unicode to support unsupported emojis, it would be far easier, more sustainable, more effective, and less burdensome on the Unicode Consortium, the public, and vendors for Unicode to just have fewer unsupported emojis.
> ...

Plain Text Alternatives

People’s goals are to have interchangeable text that includes emoji. There would be little value to having yet another format in plain text. After all, there is one already for referencing wikidata that is widely deployed:

<href a=’https://www.wikidata.org/wiki/Q29099’>

And the display format for something like what is being suggested already has a known and

well-defined display:

☄ ^Q12345^ ☄

Date/Time: Mon Nov 18 18:49:58 CST 2019
Name: James Kass
Report Type: Public Review Issue
Opt Subject: PRI #408: QID Emoji Sequences
QID Emoji represents an interesting approach to plain-text. The approach is reminiscent of suggestions made in the past to the Unicode public list which were dismissed at the time. For example, the QID material database could be just as simply referenced in plain-text by the following:

COMET + CIRCUMFLEX + Q + <the ID number in ASCII> + CIRCUMFLEX + COMET
...

Typos or wording fixes

Good catch, if we get to the point where we use that text, the corrections should be included.

Date/Time: Sun Oct 27 11:17:23 CDT 2019
Name: David Corbett
Report Type: Public Review Issue
Opt Subject: PRI #408
In “QID emoji tag sequences for flags or other symbols that represent an entity should use the QID for the flag or symbol itself if available, not the flag for the entity,” it should say “the QID of the entity” not “the flag of the entity”.

Wikidata

There are two issues: (a) Is the Unicode Consortium pushing the decision of what is emoji to Wikidata? (b) If Wikidata is a gatekeeper, are they reliable?

As for (a), the reason for using Wikidata is that it is open and growing; and new entities can be added when needed. That provides the bases for custom IDs, that can grow to suit people’s needs, but are also well-defined and discoverable. The decision as to what becomes RGI emoji still rests with the Unicode Consortium.

As for (b), Denny Vrandecic’s answer provides sufficient background for confidence in the organization. Handling of deletions is an obvious question: The relevant criterion for preventing that is “It refers to an instance of a **clearly identifiable conceptual or material entity**. The entity must be notable, in the sense that it **can be described using serious and publicly available references**.” So although deletions are rare, people should ensure that the references for anything they deploy are satisfactory. Note that if a QID were made RGI, then the Unicode Consortium would stabilize the reference as it does with other systems such as ISO country codes.

Date/Time: Mon Nov 25 10:32:13 CST 2019

Name: David Lewis
Report Type: Public Review Issue
Opt Subject: PRI #408: QID Emoji Sequences
I have to agree with others who have posted on this subject. With QID it appears that the Unicode Consortium is for some reason attempting to defeat the entire purpose of the Unicode Consortium.

The entire point of Unicode is for one body to decide for all of computing what character a particular sequence of binary digits represents across all implementations around the entire world. It does slow the process of adding new symbols considerably, but in exchange a host of issues are bypassed. If everyone implements Unicode according to the standard, there will never be any more conversion errors again. The character you expect to display is the character you WILL display, if your font supports it.
...
I don't trust a Wiki as a governing body. I don't think we should wait a couple of years for QID to get so messed up that a body of individuals have to create a QID Consortium to bring the world to one singular global QID standard.
...
Date/Time: Tue Apr 2 05:45:45 CDT 2019
Name: Andrew West
Report Type: Feedback on an Encoding Proposal
Opt Subject: Feedback on QID Emoji Proposal
...This mechanism could be seen as an attempt to deflect criticism away from the Unicode Consortium onto Wikidata and vendors, so that when the public or the press complain to the UTC that such or such an emoji is lacking, the UTC can simply shrug their shoulders and tell them to ask Wikidata to add an ID and vendors to support an emoji for that ID. Firstly, this in unfair on Wikidata, which never asked to become a repository for potential emoji, and secondly will not save the Unicode Consortium from criticism if Wikidata does not have an ID for Banana and Custard Pizza (for example) or if vendors do not support a particular ID, or if vendors implement the emoji for an ID inconsistently. The Unicode Consortium will still be seen as the people to blame, even though there is nothing the UTC can do to solve perceived issues with Wikidata and vendor implementations of Wikidata IDs.
...
Date/Time: Mon Apr 20 14:31:58 CDT 2020
Name: Denny Vrandecic
Report Type: Public Review Issue
Opt Subject: PRI 408 QID Emoji - Feedback
This is formal feedback to PRI issue #408 regarding the proposal of QID Emoji.

Wikidata is a Wikimedia project and follows the principles of open knowledge creation and curation that have led Wikipedia to be the project it is today. Wikidata's goal is to allow everyone to share in an open knowledge graph that anyone can edit and use.

Wikidata has more than 25,000 monthly contributors, and has seen more than 1.1 billion edits, creating more than 80 million Items. Each of these Items is identified by what we call a QID (short, for Q-Identifier, as the identifiers are starting with the letter Q and followed by a number). These

QIDs are meant to be quite stable: a QID can get discontinued when an Item is deleted, but the QID then never gets reused, thus not leading to ambiguity. A QID can also be forwarded to another QID when two Items are merged, but in this case the QID and their relation is recorded. Deletions happen rarely, and by definition only for Items that are not notable. The QIDs for almost all Items of wider interest have remained stable since their creation. Wikidata provides a service to resolve QIDs and get back human- and machine-readable names and descriptions of the Items of interest.

Wikidata has become a major authority hub for identity. Not because of complex processes and selective contribution requirements, but on the contrary, because of the ease of contributing and its adherence to Wikipedia's principles of openness and inclusion. Wikidata links together several thousand databases and authority files, allowing to swiftly join data indexed with ICD identifiers and Dewey Decimal Classification codes. This has led to Wikidata being described as a crystallization point of identifiers, as an authority file of authority files, or as a modern Stone of Rosetta. Even more importantly, although Wikidata only launched a few years ago, it is already being used by a growing number of institutions as an important authority file.

These institutions include, but are not limited to:
The US Library of Congress
The German National Library
Virtual International Authority File VIAF
The New York Times
Google
Museum of Modern Art
iNaturalist
Carnegie Hall
MusicBrainz
Open Street Maps
Schema.org
Quora
OCLC WorldCat
And many more.

Given that these and other authorities are already relying on and trusting Wikidata and its open processes to curating a comprehensive and current catalogue of identifiers, we are humbled and pleased to learn about the proposal to the Unicode Consortium to consider using Wikidata QIDs as an additional approach to identify the meaning of an emoji. We understand that this would allow stakeholders to expediently introduce new emojis, be able to measure their real-world adoption, and provide unambiguous and stable emoji tag sequences. We think that this is a great application of Wikidata as an identifier catalogue, and we fully support this proposal.

Lydia Pintscher, Wikimedia Deutschland, Product Manager Wikidata
Denny Vrandečić, Founder Wikidata
Joint statement

P.S.: if of interest, the Wikidata community already records a few thousand Unicode characters as being identified with a given QID. We could think that

this kind of mapping can be useful to stakeholders for example to do some form of normalization or fallback. As of the time of writing, there are 9,913 such mappings using the Property P487 (see https://w.wiki/NRB for a current list).

General Support

Date/Time: Fri Sep 27 08:54:07 CDT 2019
Name: Yannis Haralambous
Report Type: Public Review Issue
Opt Subject: QID Emojis
In my humble opinion, QID Emojis may very well become a major turning point in human communication: *for the first time billions of people will use semantically annotated entities in everyday informal communication*.
...

Identifiers, Duplicate Encoding, Confusability, Stability...

Charlotte Buff supplies some thoughtful feedback, which requires a more detailed response.

Identifiers. QID emoji should absolutely be removed from consideration for any normal kinds of identifiers (programmatic, IDNs, etc). Of course, any particular implementation could extend that, eg adding QID emoji to Hashtags, but that would only be for decorative use and not any environment where confusability is not an issue.

Duplicate Encoding. "Unicode exists to transmit information in a uniformly agreed-upon format, so there must never be two different sequences of codepoints representing the exact same concept unless that difference can be folded away through normalisation."

For letters, numbers, and normal symbols, that was certainly our goal. However, it is not always achieved even for those characters. Emoji are rather different in kind: the relation between semantics and encoding are rather looser, with a great deal of overlap. One way to think of a QID emoji is that it is part-way between a regular character and a PU (private use) character. There are strong differences from a PU character in that:

- a QID must be an emoji
    - an emoji appearance (square shape, colored independent of font color) and
    - emoji behavior (in line-break, bidi, etc.)
- its intended semantics are constrained by the associated Wikidata QID entry.

If the UTC chooses to add a QID to RGI, then at that point it would ensure that the semantics are clear, with the regular apparatus of a representative glyph, CLDR name and keywords, plus guidance if necessary.

So consider a scenario, that there is a QID for DACHSHUND that gets supported by company X. If that QID gets popular enough, and the UTC decides to make it RGI, then it gets broad support from vendors. Suppose that it doesn't get broad support, but instead just exists in a small ecosystem of products from company X, and the UTC doesn't even know it exists. The UTC might still add the DACHSHUND QID to RGI. Or the UTC might decide to encode a Unicode emoji character for

DACHSHUND (if a Lag Time option is not chosen from *Future Unicode Emoji Options (L2/21-078)*, for example.) In that case, there would be two representations of the DACHSHUND emoji. Over time, people migrate to using the RGI version, since it is far more widely supported. Given that dual representation issue is confined to emoji, and not an issue for identifiers, letters, numbers, etc., would it be a problem?

Stability. "If any object or concept with an associated QID can be represented as a tag sequence, then no such object or concept can ever be encoded as a regular emoji." Part of the process would be that the UTC would not add a regular emoji character or RGI sequence that duplicated an RGI QID emoji, or add a QID emoji to RGI that duplicated a regular emoji character or RGI sequence.

Display Fallback and Accessibility. "Screen readers would choke on QID sequences as well." The current thinking on QID would disallow multiple base characters (which matches what is in Buff's feedback). The feedback also makes a good point: that the best solution for the QBASE would be a new character, perhaps something like  (U+FFFD REPLACEMENT CHARACTER) in appearance, but with a colorful appearance. That avoids people's having to learn that an existing emoji (like ) could be the base for an unsupported QID.

It is clear that an unsupported QID would not only lack a visual appearance, but also lack a name for use in screen readers. In that, it is no more or less bad than an unsupported new emoji (which appears as a box, and has no name). That would be solved for RGI QID emoji, using the existing mechanisms. Any company that supported a non-RGI emoji should of course supply both an image *and* a name (ideally in all of the CLDR languages).

Moreover, similar to how search engines present info cards from Wikipedia, IMDb, etc. for search terms, an implementation could also choose to dynamically retrieve information about unsupported QIDs (in a privacy-preserving manner), but this would of course be optional.

Track Record. "The track record for generalised emoji mechanisms hasn't been great so far." These are good points. In most of the mentioned cases, it is non-technical issues that block wider usage. The biggest issue for RGI is combinatorial explosion. To add the red hair style to one character is doable in terms of memory usage and keyboard palette, but to add it to all the humanform emoji, in 3 genders and 5 skin-tones would be a huge cost. Similarly, subdivision flags turned out to be a long and slippery slope: adding all ~5K emoji would have vastly exceeded the goals for the number of RGI per year. On the other hand, zwj sequences and skin tones have been quite useful mechanisms, allowing us to avoid encoding thousands of characters, and provide for reasonable fallbacks on older systems.

It may well turn out that QID are not deployed widely. Because it leverages the existing TAG mechanism, the incremental effort for implementations is much smaller than for a new mechanism. And if it does succeed, even partially, it would allow companies to encode additional emoji that do not have to be approved by the Unicode Consortium. Thus companies could support a wider range of Unicode emoji than the Unicode RGI set, and interchange them with other platforms that supported the same set.

In essence, this would allow individual companies to quickly deploy and experiment with custom emoji characters, but in a manner that avoids encoding conflicts. Other companies can quickly jump on board if usage skyrockets. The resulting comparative usage data can provide strong evidence for integrating the emoji into the RGI set, making the most popular emoji available for all.

Discoverability. "To discover the intended meaning of a QID sequence in the wild, you would need to…". There is no expectation that an individual would take the time to go through the process

described in the feedback to look up an unsupported QID, any more than an individual would take the time to look up an unsupported emoji character or sequence (that appeared to them as a black box).

“Furthermore, creating colour fonts is not something the average person can easily do.” QIDs don’t magically enable an individual to have an arbitrary emoji. Realistically, there would need to be support by an OS or application for a non-RGI emoji to be useful. For those operating systems that allow for the addition of fonts and keyboard extensions, there could be a market in supplying packages that included a font and keyboard extension for a set of QIDs, but the number of individuals that would do that by themselves is not expected to be significant.

> **Date/Time:** Tue Jun 4 18:37:44 CDT 2019
> **Name:** Charlotte Buff
> **Report Type:** Public Review Issue
> **Opt Subject:** Feedback on Proposed QID Emoji Mechanism
> I wanted to inform the UTC of some critical issues concerning the proposed update to UTS #51 allowing emoji to be encoded as Wikidata QID tag sequences. I fully agree with the feedback Andrew West provided in April (cf. https://www.unicode.org/L2/L2019/19124-pubrev.html) but there are additional points he did not address.
>
> == Duplicate Encoding ==
>
> If any object or concept with an associated QID can be represented as a tag sequence, then that includes every object or concept that has already been encoded as a regular emoji.
>
> ...
> == Stability ==
>
> If any object or concept with an associated QID can be represented as a tag sequence, then no such object or concept can ever be encoded as a regular emoji.
>
> ...
>
> The QID mechanism would mean that no emoji character could ever be added to Unicode again, no ZWJ sequence could ever be approved, and no existing character could ever be emojified because QID sequences already cover all of them, or could cover them in the future. This includes the entire list of candidates for Emoji 13.
>
> == Fallback Display and Accessibility ==
>
> The fallback behaviour of QID sequences, like for all emoji tag sequences, is worthless.
>
> ...
>
> **Date/Time:** Tue Nov 5 10:23:02 CST 2019
> **Name:** Charlotte Buff

Report Type: Public Review Issue
Opt Subject: PRI #408: QID Emoji Sequences
...

As to whether QID sequences should be part of UTS #51 at all: The track record for generalised emoji mechanisms hasn't been great so far.

• Tag sequences for regional flags have been possible since 2017. In that time, the UTC has not RGI'd any new flags beyond the initial three, and only one vendor has ever decided to support a non-RGI sequence: WhatsApp with 🏴󠁵󠁳󠁴󠁸󠁿 (Flag for Texas).

...

Date/Time: Wed Jan 6 15:36:25 CST 2021
Name: asmus
Report Type: Public Review Issue
Opt Subject: Wrong closing date for PRI #408
...
PS: I fully endorse the comments by Ms. Buff. I believe the proposal to be fatally flawed for the reasons she articulates so well. It should be withdrawn with prejudice.
...
Date/Time: 2020-04-15
Name: Henri Sivonen, Mozilla (hsivonen@mozilla.com)
https://www.unicode.org/L2/L2020/20110-qid-emoji.pdf

Tangential

Most of the following feedback is tangential. The goal for QID is only for emoji, not for general purpose. And there is no point to further gradations among RGI.

Date/Time: Thu Nov 7 09:55:28 CST 2019
Date/Time: Fri Nov 8 10:28:16 CST 2019
Date/Time: Wed Nov 13 14:29:11 CST 2019
Date/Time: Mon Mar 2 11:35:30 CST 2020
Date/Time: Mon Apr 20 13:06:34 CDT 2020
Date/Time: Thu Jun 4 15:04:55 CDT 2020
Date/Time: Sat Apr 10 05:39:57 CDT 2021
Date/Time: Tue Sep 24 18:20:35 CDT 2019
Date/Time: Fri Sep 27 12:47:44 CDT 2019
Date/Time: Mon Sep 30 11:03:33 CDT 2019
Name: William Overington
Report Type: Public Review Issue
Opt Subject: Public Review 408: QID Emoji

Opt Subject: Public Review 405: Proposed Update UTS #51, Unicode Emoji