Network Working Group	J.C. Klensin
Internet-Draft	November 24, 2011
Updates: 2047, 2231 (if approved)	
Expires: May 27, 2012	

The "U" Encoding for Encoded-Words in Email

draft-klensin-encoded-word-type-u-00a

Abstract

The "Encoded Word" conventions have been used extensively in email headers and elsewhere to permit the encoding of non-ASCII characters where only ASCII ones are normally permitted. The existing specification defines only two kinds of encoding, one of which cannot be understood easily by people and the other of which has been widely discredited. This document specifies a third encoding that is easily accessible by users and much more closely tied to contemporary practices.

The current version of the proposal is intended for possible discussion in the EAI, IRI, and PRECIS WGs to see if it sheds light on other issues being discussed in those WGs. It is not, at this point, proposed for adoption.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

The "Encoded Word" conventions [RFC2047] have been used extensively in email headers and elsewhere to permit the encoding of non-ASCII characters where only ASCII ones are normally permitted. That existing encoded-word specification defines only two kinds of encoding, one of which cannot be understood easily by people ("B", the MIME "Base64" encoding) and the other of which ("Q", so-called Quoted Printable) has been widely discredited. This document specifies a third encoding, based on the "\u'NNNN'" convention, that is easily accessible by users and much more closely tied to contemporary practices.

Unlike the "B" and "Q" encodings, which were specified at a time when many coded character sets were in common use, it is now appropriate [RFC5198] to tie a new encoding specifically to Unicode [Unicode] and the corresponding ISO Standard [ISO10646], viewing conversion to local character sets, if necessary at all, to be a local matter. Consequently, this specification permits only the combination "=?iso-10646-UCS-4?u?".

If adopted, it is intended not only as an alternative to "Q" and "B", but also as an alternative to the %-encoding of Section 2.1 of the URI Specification *[RFC3986]* of UTF-8 [RFC3629] (and other) strings. %-encoding was more than adequate for its original purpose of encoding eight-bit character sets, notably ISO 8859-1 *[ISO8859-1]*, but is problematic for email (especially addresses and fields related to them) because "%" has an important historic (and still occasionally used) meaning in those contexts and because its use to encode already-encoded forms of multi-octet character sets, such as UTF-8 and Unicode, creates strings that are at least as difficult for end users to interpret as Base64.

1.1. Updated Specifications

This document, if approved, updates the Encoded-Word specification [RFC2047] and the specification for the use encoded-words with language information [RFC2231] to permit use of an additional encoding type, type "U".

1.2. Terminology

Some reasonable understanding of Encoded-Words and the Quoted-Printable, Base64, and %-encoding conventions are required to understand this introductory material but not the proposal itself.
The key words "MUST", "MUST NOT", "SHOULD", "SHOULD NOT", and "MAY" in this document are to be interpreted as defined in RFC 2119 *[RFC2119]*.

1.3. Scope and Discussion List

RFC Editor: In the unlikely event that you see this subsection, it should be removed before publication.
The current version of the proposal is intended for possible discussion in the EAI, IRI, and PRECIS WGs to see if it sheds light on other issues being discussed in those WGs. If discussions are of interest, they should occur on the mailing lists associated with those groups.
This Internet Draft is, at this point, intended only to promote discussion of a possibly-useful building block for other work. It is not proposed for adoption by the IETF for any purpose.

2. Specification

A new encoding form for encoded words is defined with code "u". The associated encoded-text string is consistent with the rules in Section 4 of RFC 2047, i.e., it consists of ASCII characters with space, tab, and "?" characters excluded. Non-ASCII characters are encoded using the \u'NNNN' form, where "NNNN" consists of four to six hexadecimal digits designating a Unicode (ISO 10646) code point. That encoding convention is defined in RFC 5137 *[RFC5137]* together with an explanation of why the quotes should be required.
As an example, the German equivalent of the string "This is nuts", would appear in the extended form of RFC 2231 (updated by verified Erratum 478 [RFC2231-Err478]) as
=?iso-10646-UCS-4+de?u?Das ist verr\u'00FC'ckt?=

3. Security Considerations

This specification does not raise any security issues that are not already present in RFC 2047 and its various updates. Because the coding is more transparent to the end user than any of Base64, Quoted Printable for non-ASCII text, or %-encoding of UTF-8, it may eliminate or reduce one possible attack vector that is present with those other approaches.

4. IANA Considerations

Because there does not appear to be a registry for either encoded-word encodings or the content-transfer-encodings on which they are based, this document requires no actions by the IANA.

5. References

5.1. Normative References

[RFC2119]	Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
[RFC2047]	Moore, K., "MIME (Multipurpose Internet Mail Extensions) Part Three: Message Header Extensions for Non-ASCII Text", RFC 2047, November 1996.
[RFC2231]	Freed, N. and K. Moore, "MIME Parameter Value and Encoded Word Extensions: Character Sets, Languages, and Continuations", RFC 2231, November 1997.
[RFC2231-Err478]	Stedfast, J., "MIME Parameter Value and Encoded Word Extensions: Character Sets, Languages, and Continuations, Erratum 478", November 2001.
[Unicode]	The Unicode Consortium. The Unicode Standard, Version 6.0.0, defined by:, "The Unicode Standard, Version 6.0.0 ", Mountain View, CA: The Unicode Consortium, 2011. ISBN 978-1-936213-01-6, 2011.

5.2. Informative References

[RFC3629]	Yergeau, F., "UTF-8, a transformation format of ISO 10646", STD 63, RFC 3629, November 2003.
[RFC3986]	Berners-Lee, T., Fielding, R. and L. Masinter, "Uniform Resource Identifier (URI): Generic Syntax", STD 66, RFC 3986, January 2005.
[RFC5137]	Klensin, J., "ASCII Escaping of Unicode Characters", BCP 137, RFC 5137, February 2008.
[RFC5198]	Klensin, J. and M. Padlipsky, "Unicode Format for Network Interchange", RFC 5198, March 2008.
[ISO8859-1]	International Organization for Standardization, "Information technology - 8-bit single byte coded graphic - character sets - Part 1: Latin alphabet No. 1", ISO Standard 8859-1:1998, 1998.
[ISO10646]	

	International Organization for Standardization, "Information Technology - Universal Multiple-octet coded Character Set (UCS)", ISO Standard 10646:2011, March 2011.

Author's Address

John C Klensin Klensin 1770 Massachusetts Ave, #322 Cambridge, MA 02140 USA Phone: +1 617 491 5735 EMail: john-ietf@jck.com