

DISPATCH Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 5, 2020

C. Bormann
Universitaet Bremen TZI
July 04, 2019

Modern Network Unicode
draft-bormann-dispatch-modern-network-unicode-00

Abstract

[RFC 5198](#) both defines common conventions for the use of Unicode in network protocols and caters for the specific requirements of the legacy protocol Telnet. In applications that do not need Telnet compatibility, some of the decisions of [RFC 5198](#) are cumbersome.

The present specification defines "Modern Network Unicode" (MNU), which is a form of [RFC 5198](#) network unicode that can be used in specifications that require the exchange of plain text over networks and where just mandating UTF-8 ([RFC 3629](#)) may not be sufficient, but there is also no desire to import all of the baggage of [RFC 5198](#).

In addition to a basic "Clean Modern Network Unicode" (CMNU), this specification defines a number of variances that can be used to tailor MNU to specific areas of application.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of [BCP 78](#) and [BCP 79](#).

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on January 5, 2020.

Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [BCP 78](#) and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

- [1.](#) Introduction [2](#)
- [1.1.](#) Terminology [2](#)
- [2.](#) Clean Modern Network Unicode [3](#)
- [3.](#) Variances [3](#)
- [3.1.](#) With lines [3](#)
- [3.2.](#) With CR-tolerant lines [3](#)
- [3.3.](#) With HT Characters [4](#)
- [3.4.](#) With CCC Characters [4](#)
- [3.5.](#) With NFKC [4](#)
- [3.6.](#) With Unicode Version NNN [4](#)
- [4.](#) Discussion [4](#)
- [5.](#) IANA considerations [5](#)
- [6.](#) Security considerations [5](#)
- [7.](#) References [5](#)
- [7.1.](#) Normative References [5](#)
- [7.2.](#) Informative References [6](#)
- Acknowledgements [6](#)
- Author's Address [6](#)

[1.](#) Introduction

(Insert copy of abstract here.)

[1.1.](#) Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [BCP 14](#) [[RFC2119](#)] [[RFC8174](#)] when, and only when, they appear in all capitals, as shown here.

Characters in this specification are named with their Unicode name notated in the usual form U+NNNN or with their ASCII names (such as CR, LF, HT, RS, NUL) [[RFC0020](#)].

2. Clean Modern Network Unicode

Clean Modern Network Unicode (CMNU) is the form of Modern Network Unicode that does not make use of any of the variances defined below. It requires conformance to [[RFC3629](#)] and [[RFC5198](#)], with the following changes:

- o Control characters (U+0000 to U+001F and U+007F to U+009F) MUST NOT be used. Note that this also excludes line endings, so a CMNU string cannot extend beyond a single line. (See also [Section 3.1](#) below.)
- o The characters U+2028 and U+2029 MUST NOT be used. (In case future Unicode versions add to the Unicode character categories Zl or Zp, any characters in these categories MUST NOT be used.)
- o Mandates of [[RFC5198](#)] that are specific to a version of Unicode are relaxed, e.g., there is no check for unassigned code points. Note that this means that a CMNU implementation may not be able to handle the normalization of a character not yet assigned in the version of Unicode that it uses. (See also [Section 3.6](#) below.)

3. Variances

In addition to CMNU, this specification describes a number of variances that can be used in the form "Modern Network Unicode with VVV", or "Modern Network Unicode with VVV, WWW, and ZZZ" for multiple variances used. Specifications that cannot directly use CMNU may be able to use MNU with one or more of these variances added.

3.1. With lines

While Clean Modern Network Unicode rules out line endings completely, line-structured text is often required. The variance "with lines" allows the use of line endings, represented by a single LF character (which is then the only control character allowed).

3.2. With CR-tolerant lines

The variance "with CR-tolerant lines" allows the sequence CR LF as well as a single LF character as a line ending. This may enable existing texts to be used as MNU without processing at the sender side (substituting that by processing at the receiver side). Note that, with this variance, a CR character cannot be used anywhere else but immediately preceding an LF character.

[3.3.](#) With HT Characters

In some cases, the use of HT characters ("TABs") cannot be completely excluded. The variance "with HT characters" allows their use, without defining their meaning (e.g., equivalence with spaces, column definitions, etc.).

[3.4.](#) With CCC Characters

Some applications of MNU may need to add specific control characters, such as RS [[RFC7464](#)] or FF characters. This variance is spelled with the ASCII name of the control character for CCC, e.g., "with RS characters".

[3.5.](#) With NFKC

Some applications require a stronger form of normalization than NFC. The variance "with NFKC" swaps out NFC and uses NFKC instead. This is probably best used in conjunction with "with Unicode version NNN".

[3.6.](#) With Unicode Version NNN

Some applications need to be sure that a certain Unicode version is used. The variance "with Unicode version NNN" (where nnn is a Unicode version number) defines the Unicode version in use as NNN. Also, it requires that only characters assigned in that Unicode version are being used.

[4.](#) Discussion

At the time of writing, RFCs are formatted in "Modern Network Unicode with CR-tolerant lines and FF characters".

The handling of line endings (not being part of CMNU, providing LF-only and LF/CRLF line endings as variances) may be controversial. In particular, calling out CR-tolerance as an extra (and often undesirable) feature may seem novel to some readers. The handling as specified here is much closer to the way line endings are handled on the software side than the cumbersome rules of [[RFC5198](#)]. More generally speaking, one could say that the present specification is intended to be used by state of the art protocols going forward, maybe less so by existing protocols that have legacy baggage.

Even in the "with CR-tolerant lines" variance, the CR character is only allowed as an embellishment of an immediately following LF character. This reflects the fact that overprinting has only seen niche usage for quite a number of decades now.

Unicode Line and Paragraph separators probably seemed like a good idea at the time, but have not taken hold. Today, their occurrence is more likely to trigger a bug or even serve as an attack.

The version-nonspecific nature of CMNU creates some fuzziness that may be undesirable but is more realistic in environments where applications choose the Unicode version with the Unicode library that happens to be available to them.

5. IANA considerations

This specification places no requirements on IANA.

6. Security considerations

The security considerations of [\[RFC5198\]](#) apply.

A variance "with NUL characters" would create specific security considerations as discussed in the security considerations of [\[RFC5198\]](#) and should therefore only be used in circumstances that do require it.

7. References

7.1. Normative References

- [RFC0020] Cerf, V., "ASCII format for network interchange", STD 80, [RFC 20](#), DOI 10.17487/RFC0020, October 1969, <<https://www.rfc-editor.org/info/rfc20>>.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", [BCP 14](#), [RFC 2119](#), DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC3629] Yergeau, F., "UTF-8, a transformation format of ISO 10646", STD 63, [RFC 3629](#), DOI 10.17487/RFC3629, November 2003, <<https://www.rfc-editor.org/info/rfc3629>>.
- [RFC5198] Klensin, J. and M. Padlipsky, "Unicode Format for Network Interchange", [RFC 5198](#), DOI 10.17487/RFC5198, March 2008, <<https://www.rfc-editor.org/info/rfc5198>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in [RFC 2119](#) Key Words", [BCP 14](#), [RFC 8174](#), DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.

7.2. Informative References

[RFC7464] Williams, N., "JavaScript Object Notation (JSON) Text Sequences", [RFC 7464](#), DOI 10.17487/RFC7464, February 2015, <<https://www.rfc-editor.org/info/rfc7464>>.

Acknowledgements

Klaus Hartke and Henk Birkholz drove the author out of his mind enough to make him finally write this up.

Author's Address

Carsten Bormann
Universitaet Bremen TZI
Postfach 330440
Bremen D-28359
Germany

Phone: +49-421-218-63921

Email: cabo@tzi.org

