Unicode Separated Values (USV)

Abstract

Unicode Separated Values (USV) is a data format that uses Unicode characters to mark parts. USV builds on ASCII separated values (ASV), and provides pragmatic ways to edit data in text editors by using visual symbols and layouts.

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 17 September 2024.

Copyright Notice

Table of Contents

1. Introduction

Unicode Separated Values (USV) is a data format useful for exchanging and converting data between various spreadsheet programs, databases, and streaming data services. This RFC explains USV.

Additionally, we propose a new media type "text/usv", to be registered with IANA.

We provide information references for a USV git repository [usv-git-repository], a programming implementation as a USV Rust crate [usv-rust-crate], and converter tools.

1.1. Requirements Language

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.

1.2. Media Type Language

The media type normative references are RFC 6838 [RFC6838], RFC 2046 [RFC2046], and RFC 4289 [RFC4289].

1.3. ABNF Language

The ABNF normative reference is RFC 5234 [RFC5234].

2. USV characters

Separators:

- File Separator (FS) is U+001C or U+241C
- Group Separator (GS) is U+001D or U+241D
- Record Separator (RS) is U+001E or U+241E
- Unit Separator (US) is U+001F or U+241F

Modifiers:

- Escape (ESC) is U+001B or U+241B
- End of Transmission (EOT) is U+0004 or U+2404

Liners:

- Carriage Return (CR) is U+000D
- Line Feed (LF) is U+000A

3. Definition of the USV Format

3.1. Data

Data comprises units, records, groups, and files.

3.2. Unit

A unit comprises content characters. It runs until a Unit Separator (US):

Example unit and unit separator:

```
<CODE BEGINS> file "unit-and-unit-separator.usv"


aaaᵁˢ


<CODE ENDS>
```

3.3. Record

A record comprises units. It runs until a Record Separator (RS):

Example record and record separator:

```
<CODE BEGINS> file "record-and-record-separator.usv"


aaaᵁˢbbbᵁˢᴿˢ


<CODE ENDS>
```

3.4. Group

A group comprises records. It runs until a Group Separator (GS):

Example group and group separator:

```
<CODE BEGINS> file "group-and-group-separator.usv"


aaaᵁˢbbbᵁˢᴿˢcccᵁˢdddᵁˢᴿˢᴳˢ


<CODE ENDS>
```

3.5. File

A file comprises groups. It runs until a file separator.

Example file and file separator:

```
<CODE BEGINS> file "file-and-file-separator.usv"

aaaᵁˢbbbᵁˢᴿˢcccᵁˢdddᵁˢᴿˢᴳˢeeeᵁˢfffᵁˢᴿˢgggᵁˢhhhᵁˢᴿˢᴳˢᶠˢ

<CODE ENDS>
```

3.6. Header

There may be an optional header appearing as the first item and with the same format as normal items. This header will contain names corresponding to the fields in the data, and should contain the same number of fields as the rest of data. The presence or absence of the header line should be indicated via the optional "header" parameter of this media type.

For example:

```
<CODE BEGINS> file "header.usv"

nameᵁˢnameᵁˢᴿˢaaaᵁˢbbbᵁˢᴿˢ

<CODE ENDS>
```

3.7. Escape (ESC)

Escape (ESC) makes the next character content.

Example: USV with a unit that contains an Escape + End of Transmission; because of the Escape, the End of Transmission is treated as content:

```
<CODE BEGINS> file "header.usv"

    aᴱˢᶜᴱᴼᵀbᵁˢ

<CODE ENDS>
```

3.8. End of Transmission (EOT)

End of Transmission (EOT) tells any reader that it can stop reading. This is can be useful for streaming data, such as to end a connection. This can also be useful for providing data files that contain USV data, then EOT, then addition non-USV information such as comments, images, attachments, etc.

- EOT tells the data reader that it can stop.

• EOT has no effect on the output content.

Example of a unit then an End of Transmission:

```
<CODE BEGINS> file "header.usv"


abcˢₛᶠₒᵢignorable


<CODE ENDS>
```

4. ABNF grammar

4.1. Semantics

usv = *files

file = *groups

group = *records

record = *units

unit = *content-characters

4.2. Syntax

usv = (header-and-body / body) '*' ; anything after the body is chaff

header-and-body = 1*unit-run / 1*record-run / 1*group-run / 1*file-run

body = *unit-run / *record-run / *group-run / *file-run

4.3. Runs

file-run = *(*liner-character file *liner-character FS)

group-run = *(*liner-character group *liner-character GS)

record-run = *(*liner-character record *liner-character RS)

unit-run = *(*liner-character unit *liner-character US)

4.4. Character classes

content-character = typical-character / ESC '*'

typical-character = '*' - special-character

special-character = US / RS / GS / FS / ESC / EOT

liner-character = CR / LF

4.5. Unicode symbols

FS = U+001C File Separator / U+241C Symbol for File Separator

GS = U+001D Group Separator / U+241D Symbol for Group Separator

RS = U+001E Record Separator / U+241E Symbol for Record Separator

US = U+001F Unit Separator / U+241F Symbol for Unit Separator

ESC = U+001B Escape / U+241B Symbol for Escape

EOT = U+0004 End of Transmission / U+2404 Symbol for End of Transmission

CR = U+000D Carriage Return

LF = U+000A Line Feed

5. Examples

5.1. Hello World

This kind of data ...

```
<CODE BEGINS> file "hello-world.txt"


hello, world


<CODE ENDS>
```

... is represented in USV as two units:

```
<CODE BEGINS> file "hello-world.usv"


hello␟world␟


<CODE ENDS>
```

If you prefer to see one unit per line, then you can add carriage returns and/or newlines:

```
<CODE BEGINS> file "hello-world-with-lines.usv"


helloᵁₛ
worldᵁₛ


<CODE ENDS>
```

5.2. Hello World Goodnight Moon

This kind of data …

```
<CODE BEGINS> file "hello-world-goodnight-moon.txt"

[ hello, world ], [ goodnight, moon ]

<CODE ENDS>
```

… is represented in USV as two records, each with two units:

```
<CODE BEGINS> file "hello-world-goodnight-moon.usv"

helloᵁₛworldᵁₛᴿₛgoodnightᵁₛmoonᵁₛᴿₛ

<CODE ENDS>
```

If you prefer to see one record per line, then you can add carriage returns and/or newlines:

```
<CODE BEGINS> file "hello-world-goodnight-moon-with-lines.usv"

helloᵁₛworldᵁₛᴿₛ
goodnightᵁₛmoonᵁₛᴿₛ


<CODE ENDS>
```

5.3. Units, Records, Groups, Files

USV with 2 units by 2 records by 2 groups by 2 files:

```
<CODE BEGINS> file "units-records-groups-files.usv"


aᵁˢbᵁˢᴿˢcᵁˢdᵁˢᴿˢᴳˢeᵁˢfᵁˢᴿˢgᵁˢhᵁˢᴿˢᴳˢᶠˢiᵁˢjᵁˢᴿˢkᵁˢlᵁˢᴿˢᴳˢmᵁˢnᵁˢᴿˢoᵁˢpᵁˢᴿˢᴳˢᶠˢ


<CODE ENDS>
```

If you prefer to see one record per line, then you can add carriage returns and/or newlines:

```
<CODE BEGINS> file "units-records-groups-files-with-lines.usv"


aᵁˢbᵁˢᴿˢ
cᵁˢdᵁˢᴿˢ
ᴳˢ
eᵁˢfᵁˢᴿˢ
gᵁˢhᵁˢᴿˢ
ᴳˢ
ᶠˢ
iᵁˢjᵁˢᴿˢ
kᵁˢlᵁˢᴿˢ
ᴳˢ
mᵁˢnᵁˢᴿˢ
oᵁˢpᵁˢᴿˢ
ᴳˢ
ᶠˢ


<CODE ENDS>
```

If you prefer to see one unit per line, then you can add carriage returns and/or newlines:

```
<CODE BEGINS> file "units-records-groups-files-with-lines.usv"


a ᵁₛ
b ᵁₛ
ᴿₛ
c ᵁₛ
d ᵁₛ
ᴿₛ
ᴳₛ
e ᵁₛ
f ᵁₛ
ᴿₛ
g ᵁₛ
h ᵁₛ
ᴿₛ
ᴳₛ
ᶠₛ
i ᵁₛ
j ᵁₛ
ᴿₛ
k ᵁₛ
l ᵁₛ
ᴿₛ
ᴳₛ
m ᵁₛ
n ᵁₛ
ᴿₛ
o ᵁₛ
p ᵁₛ
ᴿₛ
ᴳₛ
ᶠₛ


<CODE ENDS>
```

5.4. Articles

USV can format paragraphs, such as in this example data stream of articles; note the units contain leading liners and trailing liners.

```
<CODE BEGINS> file "articles.usv"


Title One
ᵁₛ
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod
tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim
veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip.
ᵁₛ ᴿₛ
Title Two
ᵁₛ
Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore
eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident,
sunt in culpa qui officia deserunt mollit anim id est laborum.
ᵁₛ ᴿₛ
Title Three
ᵁₛ
Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium
doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore
veritatis et quasi architecto beatae vitae dicta sunt explicabo.
ᵁₛ ᴿₛ


<CODE ENDS>
```

6. Source Code Examples

These source code examples demonstrate the Rust programming language and the USV Rust crate.

Units:

```
<CODE BEGINS> file "usv-rust-crate-units.rs"


use usv::*;
let str = "aᵁₛbᵁₛ";
let units: Units = str.units().collect();


<CODE ENDS>
```

Records:

```
<CODE BEGINS> file "usv-rust-crate-records.rs"


use usv::*;
let str = "a␟b␟␞c␟d␟␞";
let records: Records = str.records().collect();


<CODE ENDS>
```

Groups:

```
<CODE BEGINS> file "usv-rust-crate-groups.rs"


use usv::*;
let str = "a␟b␟␞c␟d␟␞␝e␟f␟␞g␟h␟␞␝";
let groups: Groups = str.groups().collect();


<CODE ENDS>
```

Files:

```
<CODE BEGINS> file "usv-rust-crate-groups.rs"


use usv::*;
let str = "a␟b␟␞c␟d␟␞␝e␟f␟␞g␟h␟␞␝␜i␟j␟␞k␟l␟␞␝m␟n␟␞o␟p␟␞␝␜";
let files: Files = str.files().collect();


<CODE ENDS>
```

7. MIME media type registration for text/usv

This section provides the MIME media type registration application information.

To: ietf-types@iana.org

Subject: Registration of MIME media type text/usv

MIME media type name: text

MIME subtype name: usv

Required parameters: none

7.1.   Optional parameters: charset, header

Common usage of USV is UTF-8, but other character sets defined by IANA for the "text" tree may be used in conjunction with the "charset" parameter.

The "header" parameter indicates the presence or absence of the header line. Valid values are "present" or "absent". Implementors choosing not to use this parameter must make their own decisions as to whether the header line is present or absent.

7.2.   Encoding considerations

This media type uses LF to denote line breaks. However, implementors should be aware that some implementations may not conform i.e. may incorrectly use other values.

7.3.   Security considerations

USV files contain passive text data that should not pose any risks. However, it is possible in theory that malicious binary data may be included in order to exploit potential buffer overruns in the program processing USV data. Additionally, private data may be shared via this format (which of course applies to any text data).

7.4.   Interoperability considerations

Implementors should "be conservative in what you do, be liberal in what you accept from others" (RFC 793 [8]) when processing USV data.

Implementations deciding not to use the optional "header" parameter must make their own decision as to whether the header is absent or present.

7.5.   Published specification

https://github.com/sixarm/usv

7.6.   Applications that use this media type

Spreadsheet programs, such as with import/export. Database programs, such as with loading/saving text. Data conversion utilities.

7.7.   Additional information

Magic number(s): none

File extension(s): usv

Apple macOS File Type Code(s): TEXT

Intended usage: COMMON

Author/Change controller: IESG

Contact: Joel Parker Henderson <joel@joelparkerhenderson.com>

8. IANA Considerations

We are requesting IANA to create a standard MIME media type "text/usv".

We have filed an IANA request for this, with same contact information.

9. Security Considerations

This document should not affect the security of the Internet.

10. Converters

We implement converters to/from USV and various popular data formats, including ASCII Separated Values (ASV), Comma Separated Values (CSV), JavaScript Object Notation (JSON), Microsoft Excel XML (XLSX).

- asv-to-usv[asv-to-usv-rust-crate], usv-to-asv[usv-to-asv-rust-crate]
- csv-to-usv[csv-to-usv-rust-crate], usv-to-csv[usv-to-csv-rust-crate]
- json-to-usv[json-to-usv-rust-crate], usv-to-json[usv-to-json-rust-crate]
- xlsx-to-usv[xlsx-to-usv-rust-crate], usv-to-xlsx[usv-to-xlsx-rust-crate]

The converters are provided for informational purposes. The converters are not part of the specification.

11. References

11.1. Normative References

[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>.

[RFC5234]   Crocker, D., Ed. and P. Overell, "Augmented BNF for Syntax Specifications: ABNF", STD 68, RFC 5234, DOI 10.17487/RFC5234, January 2008, <https://www.rfc-editor.org/info/rfc5234>.

[RFC6838]   Freed, N., Klensin, J., and T. Hansen, "Media Type Specifications and Registration Procedures", BCP 13, RFC 6838, DOI 10.17487/RFC6838, January 2013, <https://www.rfc-editor.org/info/rfc6838>.

[RFC2046]   Freed, N. and N. Borenstein, "Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types", RFC 2046, DOI 10.17487/RFC2046, November 1996, <https://www.rfc-editor.org/info/rfc2046>.

[RFC4289]   Freed, N. and J. Klensin, "Multipurpose Internet Mail Extensions (MIME) Part Four: Registration Procedures", BCP 13, RFC 4289, DOI 10.17487/RFC4289, December 2005, <https://www.rfc-editor.org/info/rfc4289>.

11.2. Informative References

[usv-git-repository]   Henderson, J., "USV git repository at https://github.com/sixarm/usv", 2022.

[usv-rust-crate]   Henderson, J., "USV rust crate at https://crates.io/crates/usv", 2024.

[asv-to-usv-rust-crate]   Henderson, J., "ASV to USV rust crate at https://crates.io/crates/asv-to-usv", 2024.

[usv-to-asv-rust-crate]   Henderson, J., "USV to ASV rust crate at https://crates.io/crates/usv-to-asv", 2024.

[csv-to-usv-rust-crate]   Henderson, J., "CSV to USV rust crate at https://crates.io/crates/csv-to-usv", 2024.

[usv-to-csv-rust-crate]   Henderson, J., "USV to CSV rust crate at https://crates.io/crates/usv-to-csv", 2024.

[json-to-usv-rust-crate]   Henderson, J., "JSON to USV rust crate at https://crates.io/crates/json-to-usv", 2024.

[usv-to-json-rust-crate]   Henderson, J., "USV to JSON rust crate at https://crates.io/crates/usv-to-json", 2024.

[xlsx-to-usv-rust-crate]   Henderson, J., "XLSX to USV rust crate at https://crates.io/crates/xlsx-to-usv", 2024.

[usv-to-xlsx-rust-crate]   Henderson, J., "USV to XLSX rust crate at https://crates.io/crates/usv-to-xlsx", 2024.

[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>.

Acknowledgements

Contributors

Thanks to all of the contributors.

Joel Parker Henderson
Email: joel@joelparkerhenderson.com

Author's Address

Joel Parker Henderson (EDITOR)
601 Van Ness Ave #E3-359
San Francisco, CA 94102
United States of America
Phone: 1-415-317-2700
Email: joel@joelparkerhenderson.com
URI: https://linkedin.com/in/joelparkerhenderson