AS5 Subtitle Format Draft

Rodrigo Braz Monteiro, Niels Martin Hansen, David Lamparter

Contents

1 Abstract

This document specifies the *AS5 subtitle format*, developed jointly by the Aegisub[1] and asa[2] teams in order to replace the old *Sub Station Alpha*[3] subtitle format and its extensions:

- Advanced Sub Station Alpha (ASS) implemented by VSFilter[5]
- Advanced Sub Station Alpha 2 (ASS2), also implemented by VSFilter
- Advanced Sub Station Alpha 3 (ASS3) implemented by equinox.

The goal is to create a flexible, easy to understand and powerful subtitle format that can be used in hardsubs or multiplexed into Matroska Video[7] files as softsubs.

2 File Structure

2.1 File Format

All AS5 files are *REQUIRED* to comply with the three requirements below:

- Be encoded with one of *UTF-8*[8], *UTF-16 Big Endian* [9] or *UTF-16 Little Endian* Unicode Transformation Formats. UTF-8 is preffered.
- Not to have any character below Unicode code point U+20, except for U+09, U+0A, U+0D. That is, it must be a plain-text file.
- All lines must end with Windows line endings, that is, U+0D followed by U+0A.

The character set of a subtitle file can be autodetermined by its Byte-Order Mark or by the value of the first two bytes. See below.

2.2 File Structure

The file is divided in *sections*, which are uniquely identified by a string inside square brackets, in a line of its own. From that point on, every next line is considered to be part of the last found section until another section is found. There is no end-of-section termination mark; they always end at the start of the next one or at the end of the file.

Each section is divided in lines, each line representing one command or definition. Empty lines *MUST* be ignored. It is recommended that programs generating AS5 files insert a blank line at the end of each section to increase readability. There *MUST* always be a blank line at the end of the file (as every line is required to end in a line break).

Each line in a section takes the general form of *Type: data1,data2,...,dataN*. An unknown *Type MUST* be ignored by a parser. It is recommended that subtitle editing programs keep such ignored lines in the file after re-saving it.

There are two sections which are required, *[AS5]* and *[Data]*, the equivalents of *[Script Info]* and *[Events]* in previous formats. If either of those sections is missing, the file is deemed invalid and

*(*MUST) be refused by the parser. Any other section can be ommitted from the file, and need not be implemented by all parsers. However, any unknown section *MUST* be preserved in the file by a subtitle editing program when it re-saves a file with sections that it does not recognize. It can, however, be removed at the user's discretion.

Finally, there is a special type of undefined group, *[Private:PROGNAME]*, which *MUST* be *ENTIRELY* preserved by other programs when re-saving it. This is used to store program-specific data, for example, Aegisub would create a group called *[Private:Aegisub]* to store its data inside. This type of group should be identified by the fact that it starts with *"[Private:"*.

2.2.1 [AS5]

This must be the first section in every AS5 file. If the very first line of the file is not [AS5], the file *MUST* be rejected by the parser as invalid. Note, however, that the first line is allowed to contain a Byte-Order Mark (BOM), which is the character U+FEFF encoded in the encoding used for the rest of the script[10]. The first four bytes will therefore be:

- 0xEF 0xBB 0xBF 0x5B - UTF-8 (with BOM)
- 0x5B 0x41 0x53 0x53 - UTF-8 (without BOM)
- 0xFF 0xFE 0x5B 0x00 - UTF-16 LE (with BOM)
- 0x5B 0x00 0x41 0x00 - UTF-16 LE (without BOM)
- 0xFE 0xFF 0x00 0x5B - UTF-16 BE (with BOM)
- 0x00 0x5B 0x00 0x41 - UTF-16 BE (without BOM)

It is possible, therefore, to determine the encoding of the file by checking its first two bytes.

This section *MUST* declare the following properties:

References

[1] Rodrigo Braz Monteiro, Niels Martin Hansen, David Lamparter et al., Aegisub. Application, 2005-2007.
`http://www.aegisub.net/`

[2] David Lamparter, asa. Application, 2004-2007.
`http://asa.diac24.net/`

[3] Kotus, Sub Station Alpha. Website, 1997-2003.
`http://web.archive.org/web/*/http://www.eswat.demon.co.uk/substation.html`

[4] #Anime-Fansubs, Advanced Sub Station Alpha.
`http://www.anime-fansubs.org`
`http://moodub.free.fr/video/ass-specs.doc`

[5] Gabest, VSFilter. Application, 2003-2007.
`http://sourceforge.net/projects/guliverkli/`

[6] David Lamparter, Advanced Sub Station Alpha 3. Website, 2007.
http://asa.diac24.net/ass3.pdf

[7] The Matroska project.
http://www.matroska.org/

[8] The Internet Society, RFC 3629, "'UTF-8, a transformation format of ISO 10646'". Website, 2003.
http://tools.ietf.org/html/rfc3629

[9] The Internet Society, RFC 2781, "'UTF-16, an encoding of ISO 10646'". Website, 2000.
http://tools.ietf.org/html/rfc2781

[10] Unicode, Inc, The Unicode Standard, Chapter 13. PDF, 1991-2000.
http://www.unicode.org/unicode/uni2book/ch13.pdf