Сетевая рабочая группа
Запрос для комментариев: 4180
Категория: Информационное

Y. Shafranovich
SolidMatrix Technologies, Inc.

Октябрь 2005

Общий формат и MIME-тип для файлов значений, разделенных запятыми (CSV)

Статус этого документа

В этом документе содержится информация для интернет-сообщества. Он не определяет Internet-стандарта любого рода. Распространение этого документа не ограничено.

Уведомление об авторских правах

Краткое описание

Это Рабочее предложение (RFC) документирует формат, используемый для файлов значений, разделенных запятыми (Comma-Separated Values, CSV) и регистрирует соответствующий MIME-тип "text/csv".

Содержание

1. Введение

Формат значений, разделенных запятыми (CSV) используется для обмена данными и их преобразования между различными программами работы с электронными таблицами в течение достаточно долгого времени. Удивительно, но несмотря на то, что этот формат очень распространен, он до сих пор не был официально задокументирован. Кроме того, в то время как регистрационное дерево IANA MIME включает в себя регистрацию для типа "text/tab-separated-values" ("текст/TAB-разделенные-значения"), никакие MIME-типы никогда не были зарегистрированы в IANA для CSV. В то же время, различные программы и операционные системы начали использовать различные MIME-типы для данного формата. Это Рабочее предложение (RFC) документирует формат, используемый для файлов значений, разделенных запятыми (Comma-Separated Values, CSV) и официально регистрирует соответствующий MIME-тип "text/csv", в соответствии с RFC 2048 [1].

2. Определение формата CSV

В то время как существуют различные спецификации и реализации для формата CSV (напр. [4], [5], [6] и [7]), официальной спецификации не существует, что допускает широкое разнообразие интерпретаций CSV-файлов. Этот раздел описывает формат, которому, похоже, следуют большинство реализаций:

1. Каждая запись находится на отдельной строке, ограниченной символом разрыва строки (CRLF). Например:

aaa,bbb,ccc CRLF
zzz,yyy,xxx CRLF

2. Последняя запись в файле может иметь или не иметь разрыв строки в конце. Например:

aaa,bbb,ccc CRLF
zzz,yyy,xxx

3. Допускается строка заголовка в первой строке в том же формате, что и обычная строка записи. Этот заголовок будет содержать имена, соответствующие полям в файле и должен содержать то же количество полей, что и записи в остальной части файла (наличие или отсутствие строки заголовка должно быть указано с помощью дополнительного параметра "header" ("заголовок") этого MIME-типа). Например:

field_name,field_name,field_name CRLF
aaa,bbb,ccc CRLF
zzz,yyy,xxx CRLF

4. В заголовке и каждой записи, может быть одно или несколько полей, разделенных запятыми. Каждая строка должна содержать одинаковое количество полей по всему файлу. Пробелы считаются частью поля и не должны игнорироваться. Последнее поле в записи не должно оканчиваться запятой. Например:

aaa,bbb,ccc

5. Каждое поле может быть заключено или не заключено в двойные кавычки (однако некоторые программы, такие как Microsoft Excel, не используют двойные кавычки вовсе). Если поля не заключены в двойные кавычки, то двойные кавычки не могут появляться внутри поля. Например:

"aaa","bbb","ccc" CRLF
zzz,yyy,xxx

6. Поля, содержащие разрывы строки (CRLF), двойные кавычки и запятые, должны быть заключены в двойные кавычки. Например:

"aaa","b CRLF bb","ccc" CRLF
zzz,yyy,xxx

7. Если поле заключено в двойные кавычки, то двойные кавычки появляющиеся внутри поля должны быть экранированы еще одними предшествующими двойными кавычками. Например:

"aaa","b""bb","ccc"

Грамматика ABNF [2] выглядит следующим образом:

```
file = [header CRLF] record *(CRLF record) [CRLF]
header = name *(COMMA name)
record = field *(COMMA field)
name = field
field = (escaped / non-escaped)
escaped = DQUOTE *(TEXTDATA / COMMA / CR / LF / 2DQUOTE) DQUOTE
non-escaped = *TEXTDATA
COMMA = %x2C
CR = %x0D ;как указано в разделе 6.1 RFC 2234 [2]
DQUOTE = %x22 ;как указано в разделе 6.1 RFC 2234 [2]
LF = %x0A ;как указано в разделе 6.1 RFC 2234 [2]
CRLF = CR LF ;как указано в разделе 6.1 RFC 2234 [2]
TEXTDATA = %x20-21 / %x23-2B / %x2D-7E
```

3. Регистрация MIME-типа text/csv

Этот раздел содержит заявления о регистрации медиа-типа (media-type) (в соответствии с RFC 2048 [1]).

Кому: ietf-types@iana.org	To: ietf-types@iana.org
Тема: регистрация MIME медиа-типа text/csv	Subject: Registration of MIME media type text/csv
MIME имя медиа-типа: text	MIME media type name: text
MIME имя подтипа: csv	MIME subtype name: csv
Обязательные параметры: нет	Required parameters: none
Необязательные параметры: кодировка, заголовок	Optional parameters: charset, header

В общем случае в CSV используется US-ASCII, но и другие наборы символов, определенные IANA для ветви "text" ("текст"), могут использоваться в сочетании с параметром "charset" ("кодировка").

Параметр "header" ("заголовок") указывает на наличие или отсутствие строки заголовка. Допустимые значения параметра: "present" ("присутствует") или "absent" ("отсутствует"). Разработчики, предпочитающие не использовать этот параметр, должны принимать собственные решения относительно присутствия или отсутствия строки заголовка.

Вопросы кодирования:

Как указано в разделе 4.1.1. RFC 2046 [3], этот тип носителя использует CRLF для обозначения разрыва строки. Тем не менее, разработчики должны знать, что некоторые реализации могут использовать другие значения.

Вопросы безопасности:

CSV-файлы содержат пассивные текстовые данные, которые не должны создавать каких-либо рисков. Тем не менее, теоретически возможно, что вредоносные двоичные данные могут быть включены для того, чтобы использовать потенциальные переполнения буфера в программной обработке CSV данных. Кроме того, персональные данные могут быть переданы через этот формат (что, конечно, относится к любым текстовым данным).

Вопросы совместимости:

Из-за отсутствия единой спецификации, существуют значительные различия между реализациями. Разработчики должны "быть консервативными в том, что вы делаете, быть либеральными в том, что вы принимаете от других" (RFC 793 [8]) при обработке файлов CSV. Попытку общего определения можно найти в разделе 2.

Разработчики, решившие не использовать дополнительный параметр "header" ("заголовок") должны принять собственное решение относительно того, отсутствует или присутствует заголовок.

Опубликованная спецификация:

Пока существуют многочисленные частные спецификации для различных программ и систем, нет единой "основной" спецификации для этого формата. Попытка общего определения может быть найдена в Разделе 2

Приложения, использующие этот медиа-тип:

Электронные таблицы, программы и различные утилиты преобразования данных

Дополнительная информация:

Магическое число(а): нет	Magic number(s): none
Расширение файла(ов): CSV	File extension(s): CSV
Код(ы) типа файла(ов) Macintosh: TEXT	Macintosh File Type Code(s): TEXT

Лицо и контактный адрес электронной почты для получения дополнительной информации:

Яков Шафранович ietf@shaftek.org (Yakov Shafranovich <ietf@shaftek.org>)

Назначение использования: ОБЩЕЕ	Intended usage: COMMON
Автор/Контролер изменений: IESG	Author/Change controller: IESG

4. Согласование с IANA

IANA зарегистрировал MIME-тип "text/csv" с помощью заявления, приведенного в Разделе 3 настоящего документа.

5. Вопросы безопасности

См обсуждение выше в разделе 3.

6. Благодарности

Автор хотел бы поблагодарить Дэйва Крокера (Dave Crocker), Мартин Дуэрста (Martin Duerst), Джоэл М. Халперн (Joel M. Halpern), Клайда Инграма (Clyde Ingram), Грэхам Клайн (Graham Klyne), Брюса Лилли (Bruce Lilly), Криса Лилли (Chris Lilley), и членов IESG за их полезные предложения. Особые слова благодарности идут Дэйву (Dave) за помощь с грамматикой ABNF.

Автор хотел бы также поблагодарить Хенрика Лефковеца (Henrik Lefkowetz), Маршалла Роуза (Marshall Rose), и людей из xml.resource.org за обеспечение многих инструментов, используемых для подготовки RFC и интернет-проектов.

Особая благодарность L.T.S.

7. Ссылки

7.1. Нормативные ссылки

[1] Freed, N., Klensin, J., and J. Postel, "Multipurpose Internet Mail Extensions (MIME) Part Four: Registration Procedures", BCP 13, RFC 2048, November 1996.

[2] Crocker, D. and P. Overell, "Augmented BNF for Syntax Specifications: ABNF", RFC 2234, November 1997.

[3] Freed, N. and N. Borenstein, "Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types", RFC 2046, November 1996.

7.2. Информативные ссылки

[4] Repici, J., "HOW-TO: The Comma Separated Value (CSV) File Format", 2004, <http://www.creativyst.com/Doc/Articles/CSV/CSV01.htm>.

[5] Edoceo, Inc., "CSV Standard File Format", 2004, <http://www.edoceo.com/utilis/csv-file-format.php>.

[6] Rodger, R. and O. Shanaghy, "Documentation for Ricebridge CSV Manager", February 2005, <http://www.ricebridge.com/products/csvman/reference.htm>.

[7] Raymond, E., "The Art of Unix Programming, Chapter 5", September 2003, <http://www.catb.org/~esr/writings/taoup/html/ch05s02.html>.

[8] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, September 1981.

Адрес автора

Yakov Shafranovich SolidMatrix Technologies, Inc.

EMail: ietf@shaftek.org URI: http://www.shaftek.org

Примечания переводчика:

На практике под CSV часто понимают более общий формат DSV (delimiter-separated values, значения, разделенные разделителем), который может использовать отличные от запятой разделители. Наиболее популярные в нашей стране разделители – символ табуляции и точка с запятой. Причем некоторые программы, как, например, Microsoft Excel, могут использовать тот или иной разделитель в зависимости от региональных настроек (в MS Excel – запятая в общих настройках и точка с запятой в российских).

В зависимости от разделителя форматы могут называться:

- CSV (Comma-separated values) – значения, разделенные запятыми,
- SCSV (Semi-colon-separated values) значения, разделенные точкой с запятой,
- TSV (Tab-separated values) значения, разделенные символом табуляции.

Также в качестве разделителя используются как минимум: двоеточие, пробел, вертикальная черта.

Вне зависимости от использованного разделителя, остальные рекомендации остаются неизменными.

В п.5 раздела 2 сделано не совсем верное замечание относительно использования кавычек в MS Excel (по крайней мере, начиная с версии 2003). При сохранении файла в формате CSV MS Excel в соответствии с данными указаниями предваряет двойные кавычки предшествующими, и заключает поле в двойные кавычки. Таким образом, если значение поля начинается с двойных кавычек, в начале этого поля в CSV файле, сохраненном из MS Excel, будут находиться 3 двойные кавычки.

Например, исходная строка в книге MS Excel:

1	"привет" медвед	2

Строка в сохраненном в формате CSV файле:

1,"""привет"" медвед",2

Перевод: Дмитрий Хартанович (2016г.)