Action Plan Background: CSV

Author: Carol C. H. Chou
Date: June 28, 2007

Preface:
CSV (Comma Separated Values) is the file format that is commonly used for exchanging data among different computer programs, particularly in spreadsheet or database applications. There is no single authoritative file format specification available for CSV.

1 General Description

1.1 Format Name: CSV file format

1.2 Version: Not Available

1.3 MIME media type name: text

1.4 MIME subtype: csv. RFC 4180 establishes "text/csv" as the MIME type registered with IANA [1].

1.5 Short Description: CSV is a simple file format that can be used to store tabular data separated by comma characters.

1.6 Common Extensions: .csv

1.7 Color depth: N/A

1.8 Color Space: N/A

1.9 Compression: N/A

1.10 Progressive Display: N/A

1.11 Animation: N/A

1.12 Magic number(s): N/A

1.13 Specification Requirements: With no formal documented file format specification for CSV, there exist many different CSV implementations. Most implementations comply with the following rules [1]:

1. Each column/field is separated by a comma character.
2. Each row/record is separated by a line break (CRLF). The last row may not be ended with a line break.
3. Columns containing commas, double quotes, or line breaks must be surrounded by double quote characters.
4. Spaces are considered part of column data and should not be ignored.
5. Double quotes embedded in a column must be preceded (escaped) by another double quote.

6. Each row must contain the same number of columns across the entire file. There should be no trailing comma for the last column of each row.
7. An optional header representing column names may appear at the first row. The optional header must follow the same rules as described above.

2 Essential and Distinguishing Characteristics

CSV is a special delimited data format using commas as its delimiters. It is commonly embedded with US-ASCII for recording tabular textual data. Just like plain text files, CSV files may be embedded with different text encodings such as Unicode. In addition, because there is no restriction on the type of data that CSV may contain, CSV may contain binary data as well.

2.1 Technical Metadata

Technical Metadata Element (G = general file metadata, GT = general text metadata, F = format specific metadata)	**Obligation (R = Required by spec., S= Defined in spec., D = Derived from spec., O = Optional)**
Encoding [GT]	D
Number of Rows [F]	D
Number of Columns [F]	D
Newline characters [F]	D

3 Usefulness

3.1 Version Duration: N/A

3.2 History of Prior Versions Duration: N/A

3.3 Expected Newer Versions: N/A

3.4 Existence of Publicly Available Complete Specifications: There is no formal authoritative specification for the CSV file format. RFC 4180 [1] and Repici [2] attempt to specify a common definition for CSV file format. They are the most complete documents available in describing CSV file format. The rules described in Repici [2] are mostly compatible with RFC 4180 except:

1. It allows the use of linefeed (LF), or carriage return and linefeed pair (CRLF) as line breaks (rule 2 in section 1.13).
2. It is not required to has the same number of columns across the entire file (rule 6 in section 1.13).
3. Leading and trailing spaces in the fields are ignored (rule 4 in section 1.13).

It appears that the CSV format described in Repici [2] is deduced from the CSV format used in Microsoft Excel.

3.5 Specifications-controlling Body: N/A

3.6 Related Legal Issues: There are no known legal issue for the common CSV file format that is described in Section 1.13. However, some CSV variations are proprietary and may subject to licensing or royalty.

3.7 Application and Platform Support: CSV is a simple data exchange format that is well supported in most spreadsheet and database applications [3]. Popular spreadsheet applications such as Excel, OpenOffice spreadsheet or Gnumeric all support importing and/or exporting of CSV file format. Many database applications including MySQL, Oracle, Postgres and Microsoft ACCESS also support CSV, allowing data in CSV files to be populated into database or vice versa.

There are numerous programming languages that support CSV. C++, Java and Ruby have libraries for reading, writing and manipulating CSV files. There are also many data conversion utilities for converting CSV files to/from other formats such as XML, XLS, PDF, DOC, HTML,etc. CSV file format appears to be well supported in most platforms including Windows, Mac, UNIX/Linux and even on mainframes.

3.8 Limitations: CSV does not limit the number of columns or rows that it can contain. Being designed for storing raw tabular data, CSV is not suitable for storing data manipulations, such as formula or stored procedures. CSV is limited for storing data in one worksheet/table.

3.9 Perceived Popularity:
Due to the simplicity of CSV file format, CSV has been widely adopted and became popular. As it is supported by most major spreadsheet and database software, there are still a lot of CSV files produced every year. After XML is standardized, it appears that some spreadsheet applications adopt XML-based file format, such as Open Document Spreadsheet (.ods) or Open XML Spreadsheet (.xlsx), as their data exchange format [2]. Nevertheless, most spreadsheet and database applications are still supporting CSV.

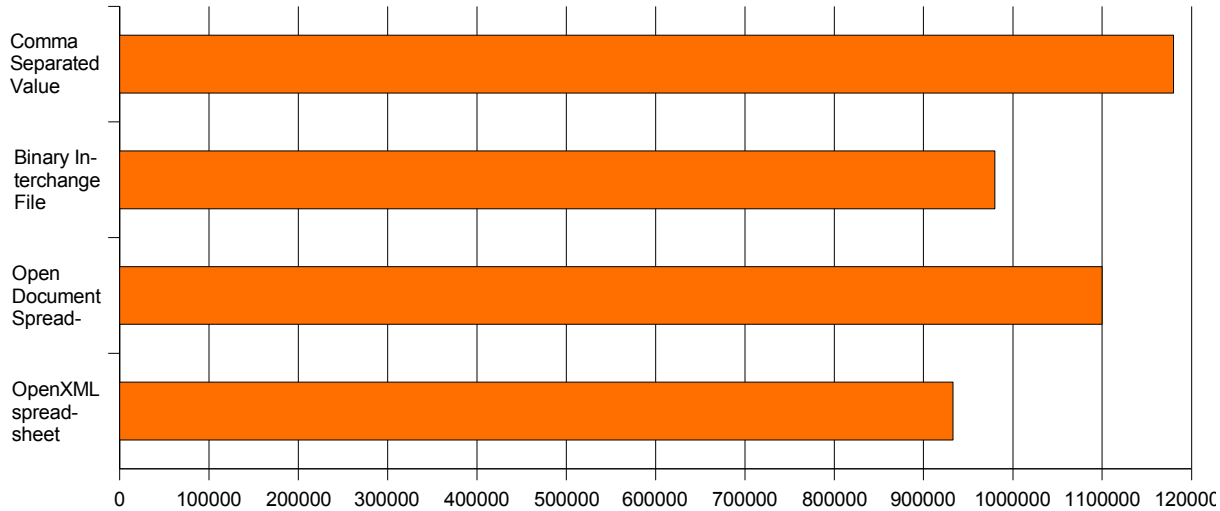

Table 1: Number of Google search hits among spreadsheet formats, performed on 6/5/2007

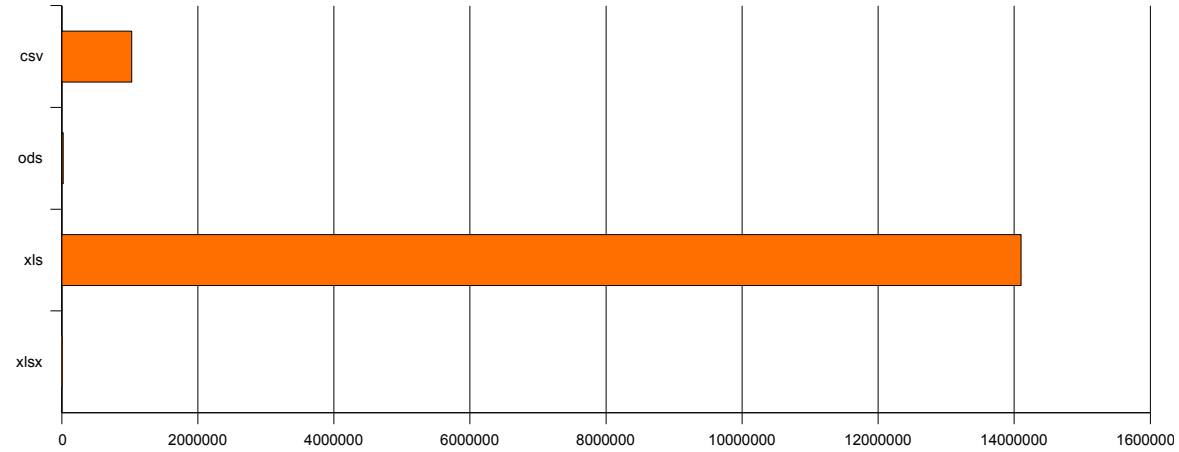

Table 2: Number of spreadsheets in Google database on June 8, 2007

The above tables compare CSV with other XML-based spreadsheet formats. Table 1 shows the number of search-hits returned from Google by using the specified keyword for each format. Table 2 shows the number of documents in the Google search database matching the specific file extensions. Although these results cannot be used as an absolute popularity comparison between CSV and other spreadsheet formats, they provide indications that there are still a lot of CSV files and information on the web. At the time of this writing, CSV does not appear to be a format approaching obsolescence.

4 Related Formats

4.1 Specification Variations:

There are many different variations of CSV file format [5][6][7]. Some variations use single quotes or apostrophes as escape characters instead of double quotes [5]. Some variations do not require each row to have the same number of columns across the entire CSV [5]. Many companies also use their own CSV variations for transferring data in their applications. Most of these variations are not documented and are not intended to be used outside of their applications [2].

5 Summary and Conclusions

CSV is a file format originally developed by Microsoft for its Excel product [4]. It is a legacy file format that has long been used as an exchange format among various applications. Its simplicity makes it one of the most supported formats among spreadsheet and database software. Unfortunately, Microsoft has never openly published the specification for CSV file format. This has led to different implementations of CSV file format. For example, some implementations only use carriages return (CR) as new lines, some use line feed (LF) and the other use linefeed plus carriage return (LFCR). Some implementations, such as Microsoft Excel, ignore leading spaces and zeros. Complexity may also arise by mixed encodings in a CSV file. For example, it is possible that a CSV file uses UTF-16 for encoding its data but uses UTF-8 for encoding its comma delimiters. Although most implementations follow the rules described in section 1.13, there exist

various implementations that can lead to different rendering or incorrect interpretation of data in CSV.

With no standard format specification and the vast variety of implementations available, preserving CSV files can be a challenge since there is no master, authoritative CSV format specification serving as the basis for format identification and validation. RFC 4180 is an attempt to specify common rules for CSV file format. Although RFC 4180 is only Informational and is not yet a standards-track specification, it is an official RFC document reviewed and approved by the Internet Engineering Steering Group (IESG). Thus , RFC 4180 appears to be the most suitable document to be used for format processing at this moment. The Florida Digital Archive will use the rules described in section 1.13, derived from RFC 4180, for performing format validation on CSV files. With no magic number and feasible well-formedness rule for format identification , CSV will tentatively be identified using the file extension.

6 References

[1] RFC 4180, "Comma Format and MIME Type for Comma-Separated Values (CSV) Files", Y. Shafranovich, SolidMatrix Technologies, Inc., October 2005

[2] "The Comma Separated Value (CSV) File Format, Create or parse data in this popular pseudo-standard format", Dominic John Repici, Creativyst, Inc.

[3] "CSV Application Support", http://en.wikipedia.org/wiki/CSV_application_support.

[4] "The CSV format", Paul Hsieh, http://www.wotsit.org/

[5] "What is a CSV file", http://www.csvreader.com/csv_format.php

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

[7] "Documentation for CSV Manager", Rice Bridge Inc., http://www.ricebridge.com/products/csvman/reference.htm

[8] "Comma-separated Values", http://en.wikipedia.org/wiki/Comma-separated_values