C CSV Parser Documentation

# Overview

The included files provide functions to parse CSV (or other character separated) files and/or strings into an accessible structure in the C programming language. The underlying parsed structures make use of doubly linked lists and are allocated on the heap of the running program.

The reason I wrote this CSV parser for C is its ease of access. While there are several other CSV parsers available, I could not find anyone that was as convenient as calling a function on a string or filename parameter and the parsed CSV structure being returned. Most of the parsers I found were more complicated than what I needed.

This CSV parser parses the input file or string and returns the parsed structure allocated on the heap.

This documentation is intended to provide information on how to use the provided CSV parser.

See Testing and Evaluation for parser testing and possible limitations.

Table of Contents

[Overview 1](#_Toc110440272)

[Basic Tutorial 2](#_Toc110440273)

[In-depth Tutorial 3](#_Toc110440274)

[Create CSV Structures 3](#_Toc110440275)

[Create CSV Cells 4](#_Toc110440276)

[Create CSV Rows 4](#_Toc110440277)

[Create CSV Tables 4](#_Toc110440278)

[Clone CSV Structures 4](#_Toc110440279)

[Free CSV Structures 4](#_Toc110440280)

[Free CSV Cells 4](#_Toc110440281)

[Free CSV Rows 5](#_Toc110440282)

[Free CSV Tables 5](#_Toc110440283)

[Print CSV Cell Structures 5](#_Toc110440284)

[Print CSV Cells 5](#_Toc110440285)

[Print CSV Rows 5](#_Toc110440286)

[Print CSV Tables 5](#_Toc110440287)

[Get CSV Structures and Coordinates 6](#_Toc110440288)

[Get CSV Structure at Specified Coordinate 6](#_Toc110440289)

[Get CSV Cell for a String 7](#_Toc110440290)

[Get Coordinates for CSV Structure 8](#_Toc110440291)

[Get Coordinates for a String 8](#_Toc110440292)

[Combine CSV Structures 8](#_Toc110440293)

[Combine Structure Pointers 8](#_Toc110440294)

[Combine Structure Clones 10](#_Toc110440295)

[Combine String into Structures 11](#_Toc110440296)

[Separating and Deleting CSV Structures 11](#_Toc110440297)

[Comparing CSV Structures 11](#_Toc110440298)

[Checking for Strings in CSV Structures 11](#_Toc110440299)

[Iterating Through Rows/Cells 11](#_Toc110440300)

[Testing and Evaluation 11](#_Toc110440301)

[Testing Process and How to Run. 11](#_Toc110440302)

[Test Results 11](#_Toc110440303)

[Evaluation and Possible Limitations 12](#_Toc110440304)

# Basic Tutorial

Once the header file and the implementation file are included in your project, CSV files and strings can be parsed using the following functions:

|  |
| --- |
| struct csv\_table \* parse\_string\_to\_csv\_table(char \* string, char delim, char quot\_char, int strip\_spaces, int discard\_empty\_cells);  struct csv\_table \* open\_and\_parse\_file\_to\_csv\_table(char \* filename, char delim, char quot\_char, int strip\_spaces, int discard\_empty\_cells); |

The functions take the string to parse (string) or the address of the file to read and parse (filename). The other function parameters control the parser behaviour:

|  |  |
| --- | --- |
| **Parameter** | **Description** |
| delim | This is the character used to separate values. For CSVs it is , but it can be set to any symbol. |
| quot\_char | Character used for making quotes, e.g. " or ' |
| strip\_spaces | Boolean flag. If true (1), parsed CSV values will be stripped of leading and trailing spaces. |
| discard\_empty\_cells | Boolean flag. If true (1), any value that is an empty string ("") will not be added to the parsed structure. |

*Note: If the delimiter is a space character* (' ')*,* strip\_spaces *and* discard\_empty\_cells *will be overridden to* 0 *(false) and* 1 *(true) respectively.*

The return value is a pointer to the structure used to store the parsed CSV values, a struct csv\_table.

struct csv\_table is the highest level of the parsed CSV structure. It is made up of a doubly linked list of struct csv\_row, which are in turn made up of a doubly linked list of struct csv\_cell. The struct csv\_cell contains the appropriate parsed string value in its str field.

Diagram

Description automatically generated

Figure 1 - Parsed CSV Structure Layout

The structure is designed to mirror the table layout in the unparsed CSV. The length field is the number of CSV rows or CSV Cells in the list of a CSV table or CSV row respectively.

Functions are provided to get a row/cell at a specific row and column index (see <some link here>).

The parsed CSV structure is allocated on the heap and therefore must be freed manually. The header files include free functions for each of the CSV structures to make sure all allocated memory is freed (see <some link here>).

# In-depth Tutorial

## Create CSV Structures

CSV structures can be created in code using the *new\_csv* functions.

### Create CSV Cells

A new CSV cell can be created using the following:

|  |
| --- |
| // Create CSV cell with no string value  struct csv\_cell \*c1 = new\_csv\_cell();  // Create a cell with a string value  struct csv\_cell \*c2 = new\_csv\_cell\_from\_str("ACell2"); |

CSV cells can have their str field populated/overwritten with a string value using populate\_csv\_cell\_str. If the str field already has a value, the function will free the current value and write in the string passed in.

|  |
| --- |
| // Populate c1 with string  populate\_csv\_cell\_str(c1, "Cell1");  // Overwrite value in c2  populate\_csv\_cell\_str(c2, "Cell2"); |

### Create CSV Rows

New CSV rows can be created with the new\_csv\_row.

|  |
| --- |
| struct csv\_row \*r1 = new\_csv\_row(); |

### Create CSV Tables

New CSV tables can be created with new\_csv\_table.

|  |
| --- |
| struct csv\_table \*t1 = new\_csv\_table(); |

## Clone CSV Structures

The clone functions are provided to clone the contents of the passed in pointer to the CSV structure.

|  |
| --- |
| struct csv\_cell \* clone\_csv\_cell(struct csv\_cell \* cell);  struct csv\_row \* clone\_csv\_row(struct csv\_row \* row);  struct csv\_table \* clone\_csv\_table(struct csv\_table \* table); |

The clone function allocates a new CSV structure on the heap and then copies the contents of the source CSV structure. Cloned rows and tables will have their child cells cloned as well.

## Free CSV Structures

CSV Structures are allocated on the heap and therefore must be freed appropriately.

### Free CSV Cells

An allocated CSV cell can be de-allocated using free\_csv\_cell.

|  |
| --- |
| free\_csv\_cell(c1); |

### Free CSV Rows

An allocated CSV row can be de-allocated using free\_csv\_row.

|  |
| --- |
| free\_csv\_row(r1); |

When a CSV row is freed, all cells in its cell list are also freed (using free\_csv\_cell). **Therefore, if you are working with a cell mapped to a row, be sure to clone the cell contents before freeing the row.**

### Free CSV Tables

An allocated CSV table can be de-allocated using free\_csv\_table.

|  |
| --- |
| free\_csv\_table(t1); |

When a CSV table is freed, all rows in its row list are also freed (using free\_csv\_row).

## Print CSV Cell Structures

The contents of csv structures can be printed using the print functions.

### Print CSV Cells

CSV cells can be printed using print\_csv\_cell.

|  |
| --- |
| print\_csv\_cell(c1); |

The function prints the string contained in the cell's str field surrounded by quotes.

|  |
| --- |
| "Cell1" |

### Print CSV Rows

CSV rows can be printed using print\_csv\_row.

|  |
| --- |
| print\_csv\_row(r1); |

The function prints all the cells in the row's cell list in the order that they appear. The cell list is surrounded by square brackets and a new line character is also printed.

|  |
| --- |
| ["Cell1", "Cell2"] |

There is also pretty\_print\_csv\_row, which prints CSV cell contents with tabulation and newlines.

|  |
| --- |
| [  "Cell1",  "Cell2"  ] |

### Print CSV Tables

A basic print can be done using the print\_csv\_table function.

|  |
| --- |
| print\_csv\_table(t1); |

This prints the set of CSV rows with their cell contents on one line. A newline character is also printed.

|  |
| --- |
| [["Cell1", "Cell2"], ["Cell3", "Cell4"]] |

The function pretty\_print\_csv\_table organizes the rows as they are printed. Each row is printed on one line with tabulation.

|  |
| --- |
| [  ["Cell1", "Cell2"],  ["Cell3", "Cell4"]  ] |

There is also super\_pretty\_print\_csv\_table, which performs a pretty print to the CSV cells in each CSV row in addition to the row organization from pretty\_print\_csv\_table.

|  |
| --- |
| [  [  "Cell1",  "Cell2"  ],  [  "Cell3",  "Cell4"  ]  ] |

## Get CSV Structures and Coordinates

### Get CSV Structure at Specified Coordinate

#### Get Structure Pointers

The following functions are provided to get a pointer to the CSV structure at a specified coordinate:

|  |
| --- |
| struct csv\_cell \* get\_cell\_ptr\_in\_csv\_row(struct csv\_row \*row, int index);  struct csv\_row \* get\_row\_ptr\_in\_csv\_table(struct csv\_table \*table, int index);  struct csv\_cell \* get\_cell\_ptr\_in\_csv\_table(struct csv\_table \*table, int rowindx, int colindx); |

The functions take the parent row/table and the index for the cell/row.

Use get\_cell\_ptr\_in\_csv\_row to get the pointer to CSV cells at a specific index.

|  |
| --- |
| // r1 = ["Cell1", "Cell2"]  // Getting "Cell1"  struct csv\_cell \*cell0 = get\_cell\_ptr\_in\_csv\_row(r1, 0);  // Getting "Cell2"  struct csv\_cell \*cell1 = get\_cell\_ptr\_in\_csv\_row(r1, 1); |

Use get\_row\_ptr\_in\_csv\_table to get the pointer to the CSV row at the specific index.

|  |
| --- |
| // t1 = [["Cell1", "Cell2"], ["Cell3", "Cell4"]]  // Getting ["Cell1", "Cell2"]  struct csv\_row \*row0 = get\_row\_ptr\_in\_csv\_table(t1, 0); |

To get a specific cell in a CSV table, get\_cell\_ptr\_in\_csv\_table can be used.

|  |
| --- |
| // t1 = [["Cell1", "Cell2"], ["Cell3", "Cell4"]]  // Getting "Cell1"  struct csv\_cell \*cell0 = get\_cell\_ptr\_in\_csv\_table(t1, 0, 0);  // Getting "Cell3"  struct csv\_cell \*cell2 = get\_cell\_ptr\_in\_csv\_table(t1, 1, 0); |

If the specified index is out of range of the parent CSV structure or the pointer to the parent CSV structure, a NULL pointer is returned.

Note: **The pointer to the CSV structure mapped in the parent structure is returned, meaning any changes made to the pointer's value will affect the parent structure. To get a deep copy, refer to Get Structure Clones.**

#### Get Structure Clones

The following functions are provided to get the clone of the CSV structure at the specified index.

|  |
| --- |
| struct csv\_cell \* get\_cell\_clone\_in\_csv\_row(struct csv\_row \*row, int index);  struct csv\_row \* get\_row\_clone\_in\_csv\_table(struct csv\_table \*table, int index);  struct csv\_cell \* get\_cell\_clone\_in\_csv\_table(struct csv\_table \*table, int rowindx, int colindx); |

In this case, the pointer returned points to a clone of the CSV structure at the specified coordinates. This means that it is separate from the parent structure it was taken from.

### Get CSV Cell for a String

The functions below are provided to get the pointer to the CSV cell in the row/table whose str field is the same string as the string parameter.

|  |
| --- |
| struct csv\_cell \* get\_cell\_for\_str\_in\_csv\_row(struct csv\_row \*row, char \* string);  struct csv\_cell \* get\_cell\_for\_str\_in\_csv\_table(struct csv\_table \*table, char \* string); |

The pointer to the CSV cell returned will be the first match found in the parent structure.

### Get Coordinates for CSV Structure

For a passed in CSV structure, the provided functions return the coordinates to the CSV structure in the passed in parent CSV structure which contains the same contents as the reference CSV structure.

|  |
| --- |
| int get\_cell\_coord\_in\_csv\_row(struct csv\_row \*row, struct csv\_cell \*cell);  int get\_row\_coord\_in\_csv\_table(struct csv\_table \*table, struct csv\_row \*row); |

For example, consider the CSV row r1 which is ["Cell1", "Cell2"]. We can get the index of the cell that contains the string "Cell1" using the following.

|  |
| --- |
| struct csv\_cell \*search\_cell = new\_csv\_cell\_from\_str("Cell1");  int index = get\_cell\_coord\_in\_csv\_row(r1, search\_cell); // index is 0 |

Note that the reference CSV structure does not have to be a part of the parent CSV structure (row/table) since the search is performed using structure content comparison.

If there is no match found, -1 is returned.

get\_cell\_coord\_in\_csv\_table is provided to get the coordinates of a CSV cell structure in the parent table with the same contents as the reference CSV structure.

|  |
| --- |
| int get\_cell\_coord\_in\_csv\_table(struct csv\_table \*table, struct csv\_cell \*cell, int \*rowindx, int \*colindx); |

Since a cell coordinate in a table has both a row and column, the function also takes pointers to store the resulting row and column index.

If a match is found, the function will populate the index pointers and return 0. Otherwise, it will return -1.

### Get Coordinates for a String

The below functions have the same behaviour as the functions discussed in Get Coordinates for CSV Structure but instead take a string parameter instead of a reference CSV structure.

|  |
| --- |
| int get\_str\_coord\_in\_csv\_row(struct csv\_row \*row, char \*string);  int get\_str\_coord\_in\_csv\_table(struct csv\_table \*table, char \*string, int \*rowindx, int \*colindx); |

In this case, the returned coordinates are to the first CSV cell that contains the same string as the reference string.

## Combine CSV Structures

The below functions add CSV structures to the list contained in the parent CSV structure.

### Combine Structure Pointers

The below functions are provided to map a child CSV structure into the list of a parent CSV structure.

|  |
| --- |
| void map\_cell\_into\_csv\_row(struct csv\_row \*rowptr, struct csv\_cell \*cellptr);  void map\_row\_into\_csv\_table(struct csv\_table \*tableptr, struct csv\_row \*rowptr); |

When a CSV structure is mapped into a parent CSV structure, the pointer to the child CSV structure is added to the end of the doubly linked list contained in the parent CSV structure. Therefore, any changes made to the child CSV structure will be reflected in the parent.

|  |
| --- |
| struct csv\_cell \*c1 = new\_csv\_cell\_from\_str("Cell1");  struct csv\_cell \*c2 = new\_csv\_cell\_from\_str("Cell2");  struct csv\_row \*r1 = new\_csv\_row();  map\_cell\_into\_csv\_row(r1, c1); // r1 = ["Cell1"]  map\_cell\_into\_csv\_row(r1, c2); // r1 = ["Cell1", "Cell2"]  populate\_csv\_cell\_str(c2, "Apple"); // c2 = "Apple"  // r1 = ["Cell1", "Apple"] |

To insert a CSV structure at a given coordinate in the parent CSV structure, the *insmap* functions can be used.

|  |
| --- |
| int insmap\_cell\_into\_csv\_row(struct csv\_row \*row, struct csv\_cell \*cell, int index);  int insmap\_row\_into\_csv\_table(struct csv\_table \*table, struct csv\_row \*row, int index);  int insmap\_cell\_into\_csv\_table(struct csv\_table \*table, struct csv\_cell \*cell, int rowindx, int colindx); |

The parameters for the function are the parent CSV structure to insert into, the CSV structure to be inserted and the coordinates at which it is inserted. As it is a mapping, the pointer is inserted into the list at the given coordinates.

|  |
| --- |
| // r1 = ["Cell1", "Apple"]  struct csv\_cell \*c3 = new\_csv\_cell\_from\_str("Cell3");  insmap\_cell\_into\_csv\_row(r1, c3, 1);  // r1 = ["Cell1", "Cell3", "Apple"] |

The return value of the function indicates if the insertion is successful. It returns 0 if it was successful and false otherwise.

The *insmap* functions can also be used to insert new mappings into the parent CSV structure. This is done by passing in the length of the parent CSV structure as the insertion index.

|  |
| --- |
| // r1 = ["Cell1", "Cell3", "Apple"]  struct csv\_cell \*c4 = new\_csv\_cell\_from\_str("Cell4");  insmap\_cell\_into\_csv\_row(r1, c4, r1->length);  // r1 = ["Cell1", "Cell3", "Apple", "Cell4"] |

Note that this is the only case where the insmap functions accept a coordinate that is outside of the range of the parent structure's list.

For insmap\_cell\_into\_csv\_table, if rowindx is table->length and colindx is 0, a new row is created and the cell is placed as the first entry.

|  |
| --- |
| // t1 = [["Cell1", "Cell2"], ["Cell3", "Cell4"]]  struct csv\_cell \*c5 = new\_csv\_cell\_from\_str("Cell5");  insmap\_cell\_into\_csv\_table(t1, c5, t1->length, 0);  // t1 = [["Cell1", "Cell2"], ["Cell3", "Cell4"], ["Cell5"]] |

If colindx is not 0 for this case, the insertion will not be done. Otherwise, rowindx is used to get the row at the appropriate index, and colindx is used as the insertion index for the retrieved row using insmap\_cell\_into\_csv\_row.

|  |
| --- |
| // t1 = [["Cell1", "Cell2"], ["Cell3", "Cell4"], ["Cell5"]]  struct csv\_cell \*c6 = new\_csv\_cell\_from\_str("Cell6");  insmap\_cell\_into\_csv\_table(t1, c6, 1, 2);  // t1 = [["Cell1", "Cell2"], ["Cell3", "Cell4", "Cell6"], ["Cell5"]] |

### Combine Structure Clones

The previous functions only map the given pointers into the parent structures, meaning that changes to the pointer will affect the parent structure. The below functions clone the structure before mapping it into the specified parent structure.

To add a structure clone to the end of the parent structure (equivalent to clone.\* and map.\*):

|  |
| --- |
| void add\_cell\_clone\_to\_csv\_row(struct csv\_row \*rowptr, struct csv\_cell \*cellptr);  void add\_row\_clone\_to\_csv\_table(struct csv\_table \*tableptr, struct csv\_row \*rowptr); |

To insert a structure at a specific coordinate in the parent structure ( equivalent to clone.\* and insmap.\*):

|  |
| --- |
| int insert\_cell\_clone\_into\_csv\_row(struct csv\_row \*rowptr, struct csv\_cell \*cellptr, int index);  int insert\_row\_clone\_into\_csv\_table(struct csv\_table \*tableptr, struct csv\_row \*rowptr, int index);  int insert\_cell\_clone\_into\_csv\_table(struct csv\_table \*tableptr, struct csv\_cell \*cellptr, int rowindx, int colindx); |

The return types are the same as the *insmap* functions.

### Combine String into Structures

The add and insertion functions are also available for strings alone. These functions create a new cell for the string and place them in the parent structure appropriately.

|  |
| --- |
| int add\_str\_to\_csv\_row(struct csv\_row \*rowptr, char \* string);  int insert\_str\_into\_csv\_row(struct csv\_row \*rowptr, char \*string, int index);  int insert\_str\_into\_csv\_table(struct csv\_table \*tableptr, char \*string, int rowindx, int colindx); |

## Separating and Deleting CSV Structures

## Comparing CSV Structures

## Checking for Strings in CSV Structures

## Iterating Through Rows/Cells

# Testing and Evaluation

## Testing Process and How to Run.

The parser was tested using the various input files in *testing/inputs*.

The input files were sourced from the GitHub repository: <https://github.com/maxogden/csv-spectrum> but also includes other added files.

The parser output was compared to the output of the parser included in the python csv module. This was done by running the C parser and python parser on the same input file, formatting the output of the python parser to match the content format printed using print\_csv\_table and performing a string comparison.

To run the test on your local system, follow the below steps.

1. Compile *testing/testparser.c* to *testing/testparser.out*.
2. Run the python script *testing/testingparser.py* in the directory *testing*.

*Note: The python script only prints the results failed test cases, to print all test cases. Use the flag* --all-tests *to print the results of all test cases. The flag* --print-files *can be used to print the contents of the input CSV files and the result.*

**Note: The starting and ending quotes are added in the printed format and are not part of the cell string.**

## Test Results

The only differing case from the python parser is for the file *inputs\my\_escaped\_quotes.csv*:

|  |
| --- |
| "ha """ha""" ha" |

The python parser returned the result:

|  |
| --- |
| [["ha "ha""" ha""]] |

The C parser returned the result:

|  |
| --- |
| [["ha "ha" ha"]] |

The C parser stripped the starting and ending quotes from the input string and also stripped 2 out of the three quotation marks within it.

This is because the parser removes quotation marks that are not escaped. To escape a quotation mark, there must be a preceding quotation mark that is not used to escape another quotation mark.

|  |
| --- |
| "a" -> a  ""a" -> a"  ""a"" -> "a"  """"a"""" -> ""a"" |

In this case, the parser counted the first two quotations as an escaped quotation and therefore added it to the final string. The third quotation was not escaped since the previous quotation is already part of an escaped pair and was therefore stripped from the final string.

To see the raw code on how quotations are stripped, see malloc\_strip\_quotes\_and\_spaces in *csvparser.c*.

## Evaluation and Possible Limitations

* Big O runtime for parsing
  + Issue with merging cells
* Big O runtime for accessing
* Big O runtime for insertions
* Quote stripping space usage and limitations