



# Tabular Data Package

A simple format for describing tabular-style data for publishing and sharing.

Author(s)	Paul Walsh, Rufus Pollock, Martin Keegan
Created	7 May 2012
Updated	2 May 2017
JSON Schema	tabular-data-package.json
Version	1

## Language

The key words <code>MUST</code>, <code>MUST</code> NOT, <code>REQUIRED</code>, <code>SHALL</code>, <code>SHALL</code> NOT, <code>SHOULD</code>, <code>SHOULD</code>, <code>NOT</code>, <code>RECOMMENDED</code>, <code>MAY</code>, and <code>OPTIONAL</code> in this document are to be interpreted as described in RFC 2119

## Introduction

Tabular Data Package is a simple container format used for publishing and sharing tabularstyle data. The format's focus is on simplicity and ease of use, especially online. In addition, the format is focused on data that can be presented in a tabular structure and in making it easy to produce (and consume) tabular data packages from spreadsheets and relational databases.

The key features of this format are the following:

- CSV (comma separated variables) for data files
- Single JSON file (datapackage.json) to describe the dataset including a schema for data files
- Reuse of existing work including other Frictionless Data specifications





### Why CSV

We chose CSV as the data format for the Tabular Data Package specification because:

- 1. CSV is very simple it is possibly the most simple data format
- 2. CSV is tabular-oriented. Most data structures are either tabular or can be transformed to a tabular structure by some form of normalization
- 3. It is open and the "standard" is well-known
- 4. It is widely supported practically every spreadsheet program, relational database and programming language in existence can handle CSV in some form or other
- 5. It is text-based and therefore amenable to manipulation and access from a wide range of standard tools (including revision control systems such as git, mercurial and subversion)
- 6. It is line-oriented which means it can be incrementally processed you do not need to read an entire file to extract a single row. For similar reasons it means that the format supports streaming.

#### More informally:

CSV is the data Kalashnikov: not pretty, but many wars have been fought with it and kids can use it.

[@pudo□ (Friedrich

Lindenberg)]

CSV is the ultimate simple, standard data format - streamable, text-based, no need for proprietary tools etc [@rufuspollock (Rufus Pollock)]

## Specification

Tabular Data Package builds directly on the Data Package specification. Thus a Tabular Data Package MUST be a Data Package and conform to the Data Package specification.

Tabular Data Package has the following requirements over and above those imposed by Data Package:





js

Lacii i esoui ce i nost DC a labulai Data Nesoui ce

## Example

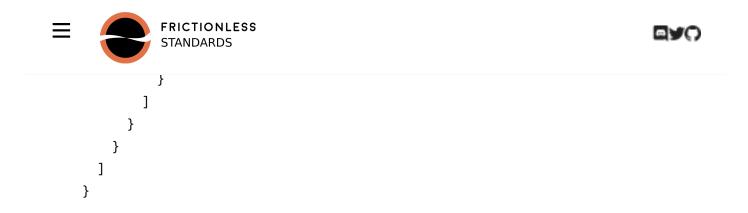
data.csv

Here's an example of a minimal tabular data package:

On disk we have 2 files:

datapackage.json

```
data.csv looks like:
  var1, var2, var3
  A,1,2.1
  B,3,4.5
datapackage.json looks like:
  {
    "profile": "tabular-data-package",
    "name": "my-dataset",
    // here we list the data files in this dataset
    "resources": [
      {
        "profile": "tabular-data-resource",
        "name": "data",
        "path": "data.csv",
        "schema": {
          "fields": [
              "name": "var1",
              "type": "string"
            },
              "name": "var2",
              "type": "integer"
            },
```



Last Updated: 7/3/2023, 9:01:36 AM