# How to produce well-formed CSV files for OpenCitations

OpenCitations processes two types of CSV files, one for metadata and one for citations. Section 1 illustrates the former's syntax, Section 2 the latter.

## 1 Metadata

Figure 1 shows an example of a well-formed CSV file containing metadata. It is structured according to a table of 11 columns, where each line corresponds to a specific document.

The 11 keys corresponding to the 11 columns are:

- **id**. The cell contains the IDs for the document described within the line. There may be one or more IDs, and they are separated by a single space (Unicode Character "SPACE", U+0020). Each ID is built as follows:

  ID abbreviation + ":" + ID value

  For example "doi:10.3233/ds-170012" indicates a DOI identifier with value "10.3233/ds-170012".

- **title**. The value corresponding to the document's title is expressed simply by a text string.

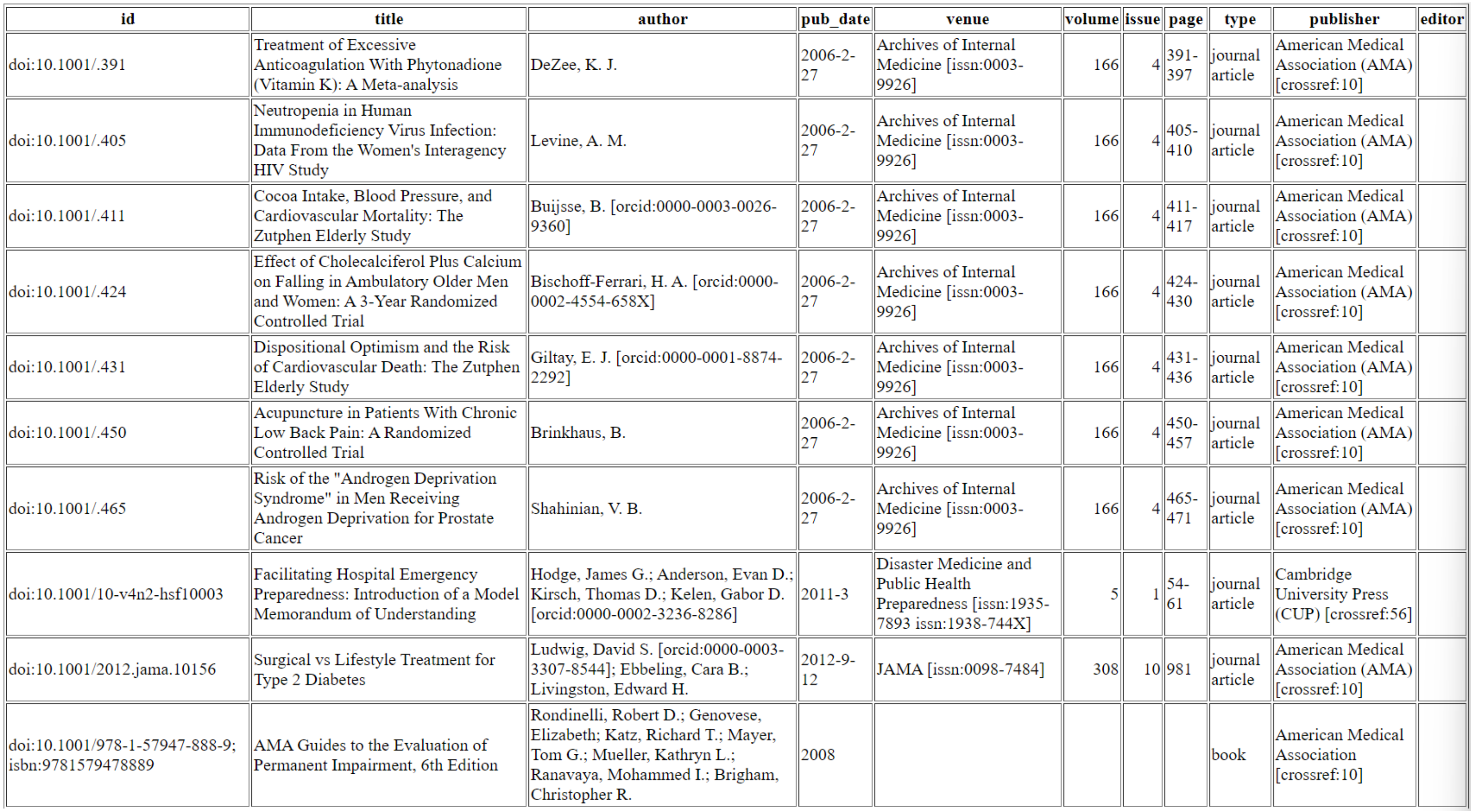
- **author**. The cell contains the data referring to the authors of the document. They are separated by a semicolon plus a single space. An author is described according to the following structure:

  Family Name + "," + " " + Given Name + " " + "[" + IDs + "]"

  The authors' IDs inside square brackets are indicated using the same structure adopted in the "id" column.

  e.g. "Peroni, Silvio [orcid:0000-0003-0530-4305]"

  If there are no IDs, there will be no square brackets either. The author's given name is not mandatory. However, the final comma will be present to indicate the incompleteness of this information (e.g. "Peroni, [orcid:0000-0003-0530- 4305]")



| id | title | author | pub_date | venue | volume | issue | page | type | publisher | editor |
|---|---|---|---|---|---|---|---|---|---|---|
| doi:10.1001/.391 | Treatment of Excessive Anticoagulation With Phytonadione (Vitamin K): A Meta-analysis | DeZee, K. J. | 2006-2-27 | Archives of Internal Medicine [issn:0003-9926] | 166 | 4 | 391-397 | journal article | American Medical Association (AMA) [crossref:10] | |
| doi:10.1001/.405 | Neutropenia in Human Immunodeficiency Virus Infection: Data From the Women's Interagency HIV Study | Levine, A. M. | 2006-2-27 | Archives of Internal Medicine [issn:0003-9926] | 166 | 4 | 405-410 | journal article | American Medical Association (AMA) [crossref:10] | |
| doi:10.1001/.411 | Cocoa Intake, Blood Pressure, and Cardiovascular Mortality: The Zutphen Elderly Study | Buijsse, B. [orcid:0000-0003-0026-9360] | 2006-2-27 | Archives of Internal Medicine [issn:0003-9926] | 166 | 4 | 411-417 | journal article | American Medical Association (AMA) [crossref:10] | |
| doi:10.1001/.424 | Effect of Cholecalciferol Plus Calcium on Falling in Ambulatory Older Men and Women: A 3-Year Randomized Controlled Trial | Bischoff-Ferrari, H. A. [orcid:0000-0002-4554-658X] | 2006-2-27 | Archives of Internal Medicine [issn:0003-9926] | 166 | 4 | 424-430 | journal article | American Medical Association (AMA) [crossref:10] | |
| doi:10.1001/.431 | Dispositional Optimism and the Risk of Cardiovascular Death: The Zutphen Elderly Study | Giltay, E. J. [orcid:0000-0001-8874-2292] | 2006-2-27 | Archives of Internal Medicine [issn:0003-9926] | 166 | 4 | 431-436 | journal article | American Medical Association (AMA) [crossref:10] | |
| doi:10.1001/.450 | Acupuncture in Patients With Chronic Low Back Pain: A Randomized Controlled Trial | Brinkhaus, B. | 2006-2-27 | Archives of Internal Medicine [issn:0003-9926] | 166 | 4 | 450-457 | journal article | American Medical Association (AMA) [crossref:10] | |
| doi:10.1001/.465 | Risk of the "Androgen Deprivation Syndrome" in Men Receiving Androgen Deprivation for Prostate Cancer | Shahinian, V. B. | 2006-2-27 | Archives of Internal Medicine [issn:0003-9926] | 166 | 4 | 465-471 | journal article | American Medical Association (AMA) [crossref:10] | |
| doi:10.1001/10-v4n2-hsf10003 | Facilitating Hospital Emergency Preparedness: Introduction of a Model Memorandum of Understanding | Hodge, James G.; Anderson, Evan D.; Kirsch, Thomas D.; Kelen, Gabor D. [orcid:0000-0002-3236-8286] | 2011-3 | Disaster Medicine and Public Health Preparedness [issn:1935-7893 issn:1938-744X] | 5 | 1 | 54-61 | journal article | Cambridge University Press (CUP) [crossref:56] | |
| doi:10.1001/2012.jama.10156 | Surgical vs Lifestyle Treatment for Type 2 Diabetes | Ludwig, David S. [orcid:0000-0003-3307-8544]; Ebbeling, Cara B.; Livingston, Edward H. | 2012-9-12 | JAMA [issn:0098-7484] | 308 | 10 | 981 | journal article | American Medical Association (AMA) [crossref:10] | |
| doi:10.1001/978-1-57947-888-9; isbn:9781579478889 | AMA Guides to the Evaluation of Permanent Impairment, 6th Edition | Rondinelli, Robert D.; Genovese, Elizabeth; Katz, Richard T.; Mayer, Tom G.; Mueller, Kathryn L.; Ranavaya, Mohammed I.; Brigham, Christopher R. | 2008 | | | | | book | American Medical Association [crossref:10] | |

Figure 1: Example of an OpenCitations' metadata CSV

- **pub_date**. This cell contains the date of publication of the document described in the row. The date is defined according to ISO 86014, the ISO standard for “Representation of dates and times”:

  YYYY-MM-DD

  YYYY indicates a four-digit year, from 0000 through 9999. MM indicates a two-digit month of the year, from 01 through 12. DD represents a two-digit day of that month, from 01 through 31. Year, month and day are separated with a hyphen “-”(Unicode Character “HYPHEN-MINUS”, U+002D), as required by the standard. It is mandatory to specify at least the publication year. On the other hand, month and day are not required. However, if the day is specified, the month must be specified.

- **venue**. The cell contains information about the venue, i.e. the bibliographical resource to which the document belongs. For example, if a row describes the metadata of a journal article, the venue will be the journal to which that article belongs. The venue is described as follows:

  Venue Title + “ ” + “[” + IDs + “]”

  The venue’s IDs inside square brackets are indicated using the same structure adopted in the “id” column. If there are no identifiers, the square brackets are not necessary.

- **volume**. This cell is only required if the entity described in the row is contained within a journal volume. The volume sequence identifier (e.g. a number) to which the entity belongs is stored here. One or more volumes constitute a journal.

- **issue**. This value is only needed if the journal article described in the row is contained within a journal issue. The issue sequence identifier (e.g. a number) to which the entity belongs is stored here. One or more issues constitute a volume of the journal.

- **page**. This key describes the page range of the resource described in the row. The value is composed of 2 numbers, first and last page respectively, divided by a hyphen “-”(Unicode Character “HYPHEN-MINUS”, U+002D).

- **type**. The string contained in this box identifies the type of resource described in the row. Here is a complete list of the currently supported bibliographic resource types: book, book chapter, book part, book section, book series, book set, book track, component, dataset (or data file), dissertation, edited book, journal, journal article, journal issue, journal volume, monograph, other, peer review, posted content (or web content), proceedings, proceedings article, proceedings series, reference book, reference entry, report, report series, standard, and standard series.

- **publisher**. This cell describes the entity responsible for making the resource available. The publisher information is structured in the following way:

  Publisher name + “ ” + “[” + IDs + “]”

  square brackets should not be entered if there is no ID.

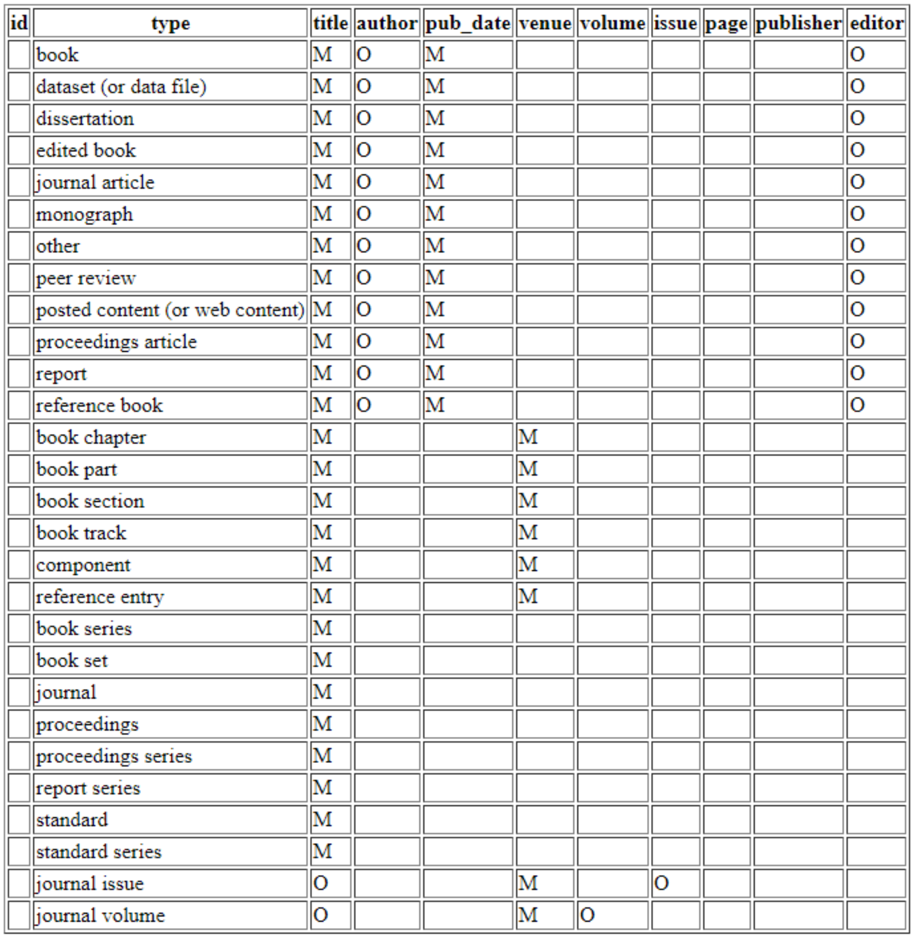
- **editor**. Since it is a human role like the author, the editor is described the same way as an author.

## 1.1 Mandatory fields

If the resource identifier is specified in the “id” field, all the other fields are optional. Conversely, if the “id” field is empty, there are mandatory fields that vary depending on the resource type:

- The fields “title”, “pub_date”, and “author” (or “editor”) are mandatory for the resources of type book, dataset (or data file), dissertation, edited book, journal article, monograph, other, peer review, posted content (or web content), proceedings article, report, and reference book. Moreover, this information is compulsory if the ”type” field is empty.
- The “title” and “venue” fields are required for the resources of type book chapter, book part, book section, book track, component, and reference entry.
- Only the “title” field is required for the resources of type book series, book set, journal, proceedings, proceedings series, report series, standard, and standard series.
- Regarding the resources of journal volume type, the fields “venue” and “volume”, or “venue” and “title”, are mandatory. Conversely, as for resources of journal issue type, the fields “venue” and “issue”, or “venue” and “title”, are mandatory.

Figure 2 summarizes the listed rules.



| id | type | title | author | pub_date | venue | volume | issue | page | publisher | editor |
|---|---|---|---|---|---|---|---|---|---|---|
| | book | M | O | M | | | | | | O |
| | dataset (or data file) | M | O | M | | | | | | O |
| | dissertation | M | O | M | | | | | | O |
| | edited book | M | O | M | | | | | | O |
| | journal article | M | O | M | | | | | | O |
| | monograph | M | O | M | | | | | | O |
| | other | M | O | M | | | | | | O |
| | peer review | M | O | M | | | | | | O |
| | posted content (or web content) | M | O | M | | | | | | O |
| | proceedings article | M | O | M | | | | | | O |
| | report | M | O | M | | | | | | O |
| | reference book | M | O | M | | | | | | O |
| | book chapter | M | | | M | | | | | |
| | book part | M | | | M | | | | | |
| | book section | M | | | M | | | | | |
| | book track | M | | | M | | | | | |
| | component | M | | | M | | | | | |
| | reference entry | M | | | M | | | | | |
| | book series | M | | | | | | | | |
| | book set | M | | | | | | | | |
| | journal | M | | | | | | | | |
| | proceedings | M | | | | | | | | |
| | proceedings series | M | | | | | | | | |
| | report series | M | | | | | | | | |
| | standard | M | | | | | | | | |
| | standard series | M | | | | | | | | |
| | journal issue | O | | | M | | O | | | |
| | journal volume | O | | | M | O | | | | |

Figure 2: Summary of mandatory fields in a metadata CSV if no identifier was specified in a specific row. “M” is an abbreviation for mandatory. Conversely, “O” stands for OR, is always present in pairs, and means that at least one element of the pair is compulsory.

## 2 Citations

Figure 3 shows an example of a well-formed CSV file containing citations. It is structured according to a table of 4 columns, where each line corresponds to a specific citation.

The 4 keys corresponding to the 4 columns are:

- **citing_id** (mandatory). This cell contains the identifier of the citing document. The identifier consists of a schema value pair, separated by a semicolon without spaces:

  ID abbreviation + ":" + ID value

  For example "pmid:23636598" indicates a PubMed identifier with value "23636598".

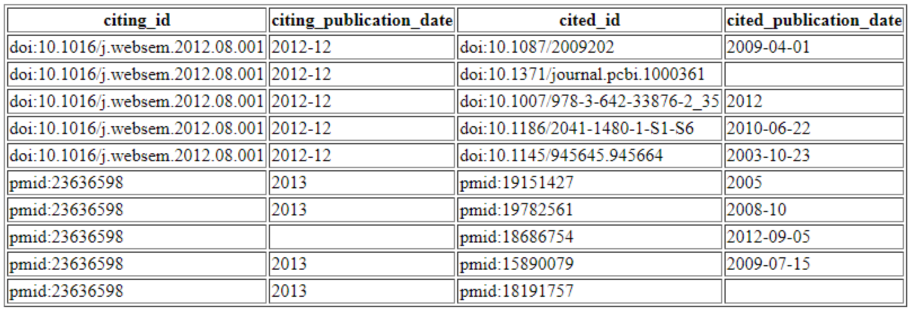
- **citing_publication_date** (optional). This cell contains the publication date of the citing document. The date is defined according to ISO 86014, the ISO standard for "Representation of dates and times":

  YYYY-MM-DD

  YYYY indicates a four-digit year, from 0000 through 9999. MM indicates a two-digit month of the year, from 01 through 12. DD represents a two-digit day of that month, from 01 through 31. Year, month and day are separated with a hyphen "-"(Unicode Character "HYPHEN-MINUS", U+002D), as required by the standard. It is mandatory to specify at least the publication year. On the other hand, month and day are not required. However, if the day is specified, the month must be specified.

- **cited_id** (mandatory). This cell contains the identifier of the cited document. It follows the same rules specified for the "citing_id" field.

- **cited_publication_date** (optional). This cell contains the publication date of the cited document. It follows the same rules specified for the "citing_publication_date" field.



| citing_id | citing_publication_date | cited_id | cited_publication_date |
|---|---|---|---|
| doi:10.1016/j.websem.2012.08.001 | 2012-12 | doi:10.1087/2009202 | 2009-04-01 |
| doi:10.1016/j.websem.2012.08.001 | 2012-12 | doi:10.1371/journal.pcbi.1000361 | |
| doi:10.1016/j.websem.2012.08.001 | 2012-12 | doi:10.1007/978-3-642-33876-2_35 | 2012 |
| doi:10.1016/j.websem.2012.08.001 | 2012-12 | doi:10.1186/2041-1480-1-S1-S6 | 2010-06-22 |
| doi:10.1016/j.websem.2012.08.001 | 2012-12 | doi:10.1145/945645.945664 | 2003-10-23 |
| pmid:23636598 | 2013 | pmid:19151427 | 2005 |
| pmid:23636598 | 2013 | pmid:19782561 | 2008-10 |
| pmid:23636598 | | pmid:18686754 | 2012-09-05 |
| pmid:23636598 | 2013 | pmid:15890079 | 2009-07-15 |
| pmid:23636598 | 2013 | pmid:18191757 | |

Figure 3: Example of a citations CSV