

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

Table 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”.

Supported ID abbreviations: “arxiv”, “doi”, “issn”, “isbn”, “jid”, “openalex”, “pmid”, “pmcid”, “wikidata”, and “wikipedia”.

- **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]”

The ID abbreviations currently supported in this cell are “orcid”, “viaf”, and “wikidata”.

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]”)

- **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: abstract, archival document, audio document, book, book chapter, book part, book section, book series, book set, computer program, data file, dataset, data management plan, dissertation, editorial, edited book, journal, journal article, journal editorial, journal issue, journal volume, newspaper, newspaper article, newspaper issue, monograph, peer review, preprint, presentation, proceedings, proceedings article, proceedings series, reference book, reference entry, report, report series, retraction notice, standard, series, standard series, and web content.
- **publisher.** This cell describes the entity responsible for making the resource available. The publisher information is structured in the following way:

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

The ID abbreviations currently supported in this cell are “crossref” and “ror”.

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.

Table 1: A sample of ten documents characterized by their corresponding metadata attributes

id	title	author	pub date	venue	volume	issue	page	type	publisher	editor
doi:10.1007/978-403-006628-8_8	The SPAR Ontologies	Peron, Silvio [orcid:0009-0003-4530-3205], Shotton, David [orcid:0000-0001-5006-523X]	2018	17th ISWC [doi:10.1007/978-403-006628-8_8]			119-136	book chapter	Springer International Publishing	[crossref:227]
doi:10.3233/DS-770112	Automating semantic publishing	Peron, Silvio [orcid:0009-0034-0530-4305]	2017	Data Science [issn:2451-8484; issn:2451-8412]	1	1-2	155-173	journal article	IOS Press	[crossref:237]
doi:10.1007/978-3-16-00160-3	Literatur		2005					book	Springer Science and Business Media LLC	[crossref:27]
isbn:9783709211444	New Waves in Philosophy of Law		2011					book	Springer Science and Business Media LLC	[crossref:27]
isbn:9780230270004	Governing Savage	Markus, Andrew	2020-07-31					book	Mar. Maksymilian Del Informa UK Limited	[crossref:301]
isbn:9780230316045	News worthy	Borbas, Samantha	2020-06-24					book	Walter de Gruyter GmbH	[crossref:374]
isbn:9780230316045	On the theory of convection of electrons in metals	Gladkov, S. O.	2017-05	High Temperature [issn:0018-151X; issn:1608-3156]	55	3	321-325	journal article	Pleiades Publishing Ltd	[crossref:317]
doi:10.4234/ijdmst.20115830	Stability of boiling shock	Aveecov, A. A.	2017-09	High Temperature [issn:0018-151X; issn:1608-3156]	55	5	732-760	journal article	Pleiades Publishing Ltd	[crossref:327]
doi:10.1134/s0915151x17050029	The high-temperature effect on concrete and radioactive effect on concrete	Zhukhin, A. I.	2017-09	High Temperature [issn:0018-151X; issn:1608-3156]	55	5	767-776	journal article	Pleiades Publishing Ltd	[crossref:327]
doi:10.1134/s0915151x18010169	Relaxation Test for and Lorient Gases in Shock Waves	Skrebkov, O. V.	2018-01	High Temperature [issn:0018-151X; issn:1608-3156]	56	1	77-83	journal article	Pleiades Publishing Ltd	[crossref:317]

1.1 Mandatory fields

If there are one or more ids and the volume or the issue is specified, then it is mandatory also to specify the venue and type, which must be one of “journal article”, “journal volume” or “journal issue”. In all other cases, the presence of one or more ids makes all other fields 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 abstract, archival document, audio document, book, computer program, data management plan, dataset (or data file), dissertation, edited book, editorial, journal article, journal editorial, monograph, newspaper article, other, peer review, posted content (or web content), preprint, presentation, proceedings article, reference book, report, and retraction notice. Moreover, this information is compulsory if the “type” field is empty.
- Only the “title” field is required for the resources of type book series, book set, journal, newspaper, 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 and newspaper issue type, the fields “venue” and “issue”, or “venue” and “title”, are mandatory.

Table 2 summarizes the listed rules.

2 Citations

Table 4 shows an example of a well-formed CSV file containing citations. It is structured according to a table of 2 columns, where each line corresponds to a specific citation.

The 2 keys corresponding to the 2 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”.

Supported ID abbreviations: “arxiv”, “doi”, “issn”, “isbn”, “jid”, “openalex”, “pmid”, “pmcid”, “wikidata”, and “wikipedia”.

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

Table 2: Summary of mandatory fields in a metadata CSV. “M” is an abbreviation for mandatory. “OR” is always present in pairs and means that at least one element of the pair is compulsory.

id	type	title	author	pub_date	venue	volume	issue	page	publisher	editor
	abstract	M	OR	M						OR
	archival document	M	OR	M						OR
	audio document	M	OR	M						OR
	book	M	OR	M						OR
	computer program	M	OR	M						OR
	data management plan	M	OR	M						OR
	dataset (or data file)	M	OR	M						OR
	dissertation	M	OR	M						OR
	edited book	M	OR	M						OR
	editorial	M	OR	M						OR
	journal article	M	OR	M						OR
	journal editorial	M	OR	M						OR
	monograph	M	OR	M						OR
	newspaper article	M	OR	M						OR
	other	M	OR	M						OR
	peer review	M	OR	M						OR
	posted content (or web content)	M	OR	M						OR
	preprint	M	OR	M						OR
	presentation	M	OR	M						OR
	proceedings article	M	OR	M						OR
	reference book	M	OR	M						OR
	report	M	OR	M						OR
	retraction notice	M	OR	M						OR
	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								
	newspaper	M								
	proceedings	M								
	proceedings series	M								
	report series	M								
	standard	M								
	standard series	M								
	journal issue	OR		M			OR			
	newspaper issue	OR		M			OR			
	journal volume	OR		M		OR				

Table 3: Additional mandatory fields when id is not specified. If volume or issue is specified for these resource types, then venue must also be mandatory.

Resource type	If volume specified	If issue specified
editorial	venue required	venue required
journal editorial	venue required	venue required
journal article	venue required	venue required
journal volume	venue required	-
journal issue	-	venue required
newspaper article	-	venue required
newspaper issue	-	venue required

Table 4: A sample of ten citations characterized by their related attributes

citing_id	cited_id
doi:10.1016/j.websem.2012.08.001	doi:10.1087/2009202
doi:10.1016/j.websem.2012.08.001	doi:10.1371/journal.pcbi.1000361
doi:10.1016/j.websem.2012.08.001	doi:10.1007/978-3-642-33876-2_35
doi:10.1016/j.websem.2012.08.001	doi:10.1186/2041-1480-1-S1-S6
doi:10.1016/j.websem.2012.08.001	doi:10.1145/945645.945664
pmid:23636598	pmid:19151427
pmid:23636598	pmid:19782561
pmid:23636598	pmid:18686754
pmid:23636598	pmid:15890079
pmid:23636598	pmid:18191757