# Manubot usage guidelines

This repository uses [Manubot](https://manubot.org) to automatically produce a manuscript from the source in the <content> directory. Check out the [Manubot catalog](https://manubot.org/catalog/) for examples of what is possible when writing with Manubot. Try editing the [demo manuscript](https://github.com/manubot/try-manubot) to quickly test Manubot formatting and citations.

## Manubot markdown

Manuscript text should be written in markdown files located in the <content> directory. Markdown files are identified by their .md extension and ordered according to their two-digit prefix (e.g. 01., 02., … 99.).

For basic formatting, check out the [CommonMark Help](https://commonmark.org/help/) page for an introduction to the formatting options provided by standard markdown. In addition, Manubot supports an extended version of markdown, tailored for scholarly writing, which includes [Pandoc’s Markdown](https://pandoc.org/MANUAL.html#pandocs-markdown) and the extensions discussed below.

The content/02.delete-me.md file in the Rootstock repository shows many of the elements and formatting options supported by Manubot. See the [raw markdown](https://gitlab.com/manubot/rootstock/blob/main/content/02.delete-me.md#L) in this file and compare it to the [rendered manuscript](https://manubot.github.io/rootstock/).

Within a paragraph in markdown, single newlines are interpreted as whitespace (same as a space). A paragraph’s source does not need to contain newlines. However, “one paragraph per line” makes the git diff less precise, leading to less granular review commenting, and makes conflicts more likely. Therefore, we recommend using [semantic linefeeds](https://rhodesmill.org/brandon/2012/one-sentence-per-line/) — newlines between sentences. We have found that “one sentence per line” is preferable to “word wrap” or “one paragraph per line”.

### Tables

Manubot supports [markdown tables](https://help.github.com/articles/organizing-information-with-tables/).

| Column 1 | Column 2 | Column 3 |  
|----------|----------|----------|  
| value\_a | 1 | 47 |  
| value\_b | 2 | 56 |  
  
Table: Caption for this example table. {#tbl:example-id}

Support for table numbering and citation is provided by [pandoc-tablenos](https://github.com/tomduck/pandoc-tablenos). Above, {#tbl:example-id} sets the table ID, which creates an HTML anchor and allows citing the table like @tbl:example-id. For easy creation of markdown tables, check out the [Tables Generator](https://www.tablesgenerator.com/markdown_tables) webapp.

### Figures

Figures can be included with the following markdown:

![Caption for the example figure.](url\_or\_path\_to\_figure){#fig:example-id}

The blank line before the figure is required. Support for figure numbering and citation is provided by [pandoc-fignos](https://github.com/tomduck/pandoc-fignos). This figure can be cited in the text using @fig:example-id. In context, a figure citation may look like: Figure {@fig:example-id}B shows ….

For images created by the manuscript authors that are hosted elsewhere on GitHub, we recommend using a [versioned](https://help.github.com/articles/getting-permanent-links-to-files/) GitHub URL to embed figures, thereby preserving exact image provenance. When embedding SVG images hosted on GitHub, it’s necessary to append ?sanitize=true to the raw.githubusercontent.com URL. For example:

https://raw.githubusercontent.com/greenelab/scihub/572d6947cb958e797d1a07fdb273157ad9154273/figure/citescore.svg?sanitize=true

Figures placed in the <content/images> directory can be embedded using their relative path. For example, we embed an [ORCID](https://orcid.org/) icon inline using:

![ORCID icon](images/orcid.svg){height="13px"}

The bracketed text following the image declaration is interpreted by Pandoc’s [link\_attributes](https://pandoc.org/MANUAL.html#extension-link_attributes) extension. For example, the following will override the figure number to be “S1” and set the image width to 5 inches:

{#fig:supplement tag="S1" width="5in"}

We recommend always specifying the width of SVG images (even if just width="100%"), since otherwise SVGs may not render properly in the [WeasyPrint](https://weasyprint.org/) PDF export.

### Citations

Manubot supports [Pandoc citations](https://pandoc.org/MANUAL.html#citations), but with added support for citing persistent identifiers directly. Citations are processed in 3 stages:

1. Pandoc parses the input Markdown to locate citation keys.
2. The [pandoc-manubot-cite filter](https://github.com/manubot/manubot#pandoc-filter) automatically retrieves the bibliographic metadata for citation keys.
3. The [pandoc-citeproc filter](https://github.com/jgm/pandoc-citeproc/blob/master/man/pandoc-citeproc.1.md) renders in-text citations and generates styled references.

When citing persistent identifiers, citation keys should be formatted like @prefix:accession, where prefix is one of the options described below. When choosing which source to use for a citation, we recommend the following order:

1. DOI (Digital Object Identifier), cite like @doi:10.15363/thinklab.4. Shortened versions of DOIs can be created at [shortdoi.org](http://shortdoi.org/). shortDOIs begin with 10/ rather than 10. and can also be cited. For example, Manubot will expand @doi:10/993 to the DOI above. We suggest using shortDOIs to cite DOIs containing forbidden characters, such as ( or ).
2. PubMed Central ID, cite like @pmc:PMC4497619.
3. PubMed ID, cite like @pubmed:26158728.
4. *arXiv* ID, cite like @arxiv:1508.06576v2.
5. ISBN (International Standard Book Number), cite like @isbn:9781339919881.
6. URL / webpage, cite like @https://nyti.ms/1QUgAt1. URL citations can be helpful if the above methods return incorrect metadata. For example, @doi:10.1038/ng.3834 [incorrectly handles](https://github.com/manubot/manubot/issues/158) the consortium name resulting in a blank author, while @https://doi.org/10.1038/ng.3834 succeeds. Similarly, @https://doi.org/10.1101/142760 is a [workaround](https://github.com/manubot/manubot/issues/16) to set the journal name of bioRxiv preprints to *bioRxiv*.
7. Wikidata Items, cite like @wikidata:Q50051684. Note that anyone can edit or add records on [Wikidata](https://www.wikidata.org), so users are encouraged to contribute metadata for hard-to-cite works to Wikidata.
8. Any other compact identifier supported by <https://bioregistry.io>. Manubot uses the Bioregistry to support [hundreds](https://github.com/manubot/manubot/blob/a4af68125ab6bb322fdfb1689cfeec09f0cb1b60/manubot/cite/handlers.py#L155-L1485) of [prefixes](https://bioregistry.io/registry/). For example, citing @clinicaltrials:NCT04280705 will produce the same bibliographic metadata as @https://bioregistry.io/clinicaltrials:NCT04280705 or @https://clinicaltrials.gov/ct2/show/NCT04280705.
9. For references that do not have any of the above persistent identifiers, the citation key does not need to include a prefix. Citing @old-manuscript will work, but only if reference metadata is [provided manually](#reference-metadata).

Manubot is able to infer certain prefixes, such some citations can be formatted like @accession (without a prefix). Examples includes DOIs like @10.15363/thinklab.4 or @10/993, PMC / PubMed identifiers like @PMC4497619 or @26158728, arXiv identifier like @1508.06576v2, and Wikidata identifiers like @Q50051684. To disable citekey prefix inference, add the following to metadata.yaml:

pandoc:  
 manubot-infer-citekey-prefixes: false

Cite multiple items at once like:

Here is a sentence with several citations [@doi:10.15363/thinklab.4; @pubmed:26158728; @arxiv:1508.06576; @isbn:9780394603988].

Note that multiple citations must be semicolon separated. Be careful not to cite the same study using identifiers from multiple sources. For example, the following citations all refer to the same study, but will be treated as separate references: [@doi:10.7717/peerj.705; @pmc:PMC4304851; @pubmed:25648772].

The citation key syntax is described in the [Pandoc manual](https://pandoc.org/MANUAL.html#citation-syntax):

Unless a citation key start with a letter, digit, or \_, and contains only alphanumerics and internal punctuation characters (:.#$%&-+?<>~/), it must be surrounded by curly braces, which are not considered part of the key. In @Foo\_bar.baz., the key is Foo\_bar.baz. The final period is not *internal* punctuation, so it is not included in the key. In @{Foo\_bar.baz.}, the key is Foo\_bar.baz., including the final period. The curly braces are recommended if you use URLs as keys: [@{https://example.com/bib?name=foobar&date=2000}, p. 33].

If a citation key does not fully match [this online regex](https://regex101.com/r/mXZyY2/latest) (for example, contains characters such as ; or = or end with a non-alphanumeric character such as /), make sure to surround it with curly braces or use the [citation aliases](#citation-aliases) workaround below.

Prior to Rootstock commit [6636b91](https://github.com/manubot/rootstock/commit/6636b912c6b41593acd2041d34cd4158c1b317fb) on 2020-01-14, Manubot processed citations separately from Pandoc. Switching to a Pandoc filter improved reliability on complex documents, but restricted the syntax of citation keys slightly. Therefore, users upgrading Rootstock may find some citations become invalid. By default, pandoc-manubot-cite does not fail upon invalid citations, although this can be changed by adding the following to metadata.yaml:

pandoc:  
 manubot-fail-on-errors: true

#### Citation aliases

The system also supports citation aliases, which map from one citation key (the “alias” or “tag”) to another. Aliases are recommended for the following applications:

1. A citation key contains forbidden characters.
2. A single reference is cited many times. Therefore, it might make sense to define an alias, so if the citation updates (e.g. a newer version becomes available), only a single change is required.

Aliases can be defined using Markdown’s [link reference syntax](https://spec.commonmark.org/0.29/#link-reference-definitions) as follows:

Citing a URL containing a `?` character [@my-url].  
Citing a DOI containing parentheses [@my-doi].  
  
[@my-url]: https://openreview.net/forum?id=HkwoSDPgg  
[@my-doi]: doi:10.1016/S0022-2836(05)80360-2

This syntax is also used by [pandoc-url2cite](https://github.com/phiresky/pandoc-url2cite). Make sure to place these link reference definitions in their own paragraphs. These paragraphs can be in any of the content Markdown files.

Another method for defining aliases is to define pandoc.citekey-aliases in metadata.yaml:

pandoc:  
 citekey-aliases:  
 my-url: https://openreview.net/forum?id=HkwoSDPgg  
 my-doi: doi:10.1016/S0022-2836(05)80360-2

## Reference metadata

Manubot stores the bibliographic details for references (the set of all cited works) as CSL JSON ([Citation Style Language Items](http://citeproc-js.readthedocs.io/en/latest/csl-json/markup.html#csl-json-items)). Manubot automatically generates CSL JSON for most persistent identifiers (as described in [Citations](#citations) above). In some cases, automatic metadata retrieval fails or provides incorrect or incomplete information. Errors are most common for references generated from scraping HTML metadata from websites. This occurs most frequently for https/http/url citations as well as Bioregistry prefixes without explicit support listed above. Therefore, Manubot supports user-provided metadata, which we refer to as “manual references”. When a manual reference is provided, Manubot uses the supplied metadata and does not attempt to generate it.

Manubot searches the content directory for files that match the glob pattern manual-references\*.\* and expects that these files contain manual references. <content/manual-references.json> is the default file to specify custom CSL JSON metadata. Manual references are matched to citations using their “id” field. For example, to manually specify the metadata for the citation @https://github.com/manubot/rootstock, add a CSL JSON Item to manual-references.json that contains the following excerpt:

"id": "https://github.com/manubot/rootstock",

The metadata for unhandled citations — any citation key that is a not a supported persistent ID — must be provided in a manual reference file (e.g. manual-references.json) or an error will occur. For example, to cite @private-message in a manuscript, a corresponding CSL JSON Item is required, such as:

{  
 "id": "private-message",  
 "type": "personal\_communication",  
 "title": "Personal communication with Doctor X"  
}

All manual references must provide values for the “id” and “type” fields. For guidance on what CSL JSON should be like for different document types, refer to [these examples](https://github.com/aurimasv/zotero-import-export-formats/blob/a51c342e66bebd97b73a7230047b801c8f7bb690/CSL%20JSON.json).

Manubot offers some support for other bibliographic metadata formats besides CSL JSON, by delegating conversion to the pandoc-citeproc --bib2json [utility](https://github.com/jgm/pandoc-citeproc/blob/master/man/pandoc-citeproc.1.md#convert-mode). Formats are inferred from filename extensions. So, for example, to provide metadata for @https://github.com/manubot/rootstock in BibTeX format, create the file content/manual-references.bib and create an item whose definition starts with the excerpt:

@misc{https://github.com/manubot/rootstock,

Processed reference metadata in CSL JSON format, either generated by Manubot or specified via manual references, is exported to references.json. This file is located in the output branch on GitHub or in the output subdirectory of local builds. The “id” field in references.json and in the final manuscript uses a shortened ID that is derived from the original ID. For debugging information, see citations.tsv, which shows citation identifiers as they progress through the processing pipeline. In order to freeze all references, rather than have Manubot regenerate them during future builds, copy the references.json output file to content with a filename matching the manual-references\*.json pattern. One tip is to embed the date references.json was generated into the frozen manual reference filename, like content/manual-references-2019-06-21.json.

## Manuscript metadata

<content/metadata.yaml> contains manuscript metadata that gets passed through to Pandoc, via a [yaml\_metadata\_block](https://pandoc.org/MANUAL.html#extension-yaml_metadata_block). metadata.yaml should contain the manuscript title, authors list, keywords, and lang ([language tag](https://www.w3.org/International/articles/language-tags/)).

When the date field is missing or null, Manubot uses the current time for the publication date. This is ideal for manuscripts that are being actively written, but once complete it might make sense to set an explicit date (ISO-format like ‘2022-10-31’), such that future minor changes do not update the publication date. The generated date will still reflect the time of the Manubot build. Manubot uses the [timezone](https://en.wikipedia.org/wiki/List_of_tz_database_time_zones) specified in [build.sh](build/build.sh) for the generated date. For example, setting the TZ environment variable to Etc/UTC directs the Manubot to use Coordinated Universal Time.

We recommend authors add themselves to metadata.yaml via pull request (when requested by a maintainer), thereby signaling that they’ve read and approved the manuscript. The following YAML shows the supported key–value pairs for an author:

github: dhimmel # strongly suggested  
name: Daniel S. Himmelstein # mandatory  
initials: DSH # optional  
orcid: 0000-0002-3012-7446 # mandatory  
twitter: dhimmel # optional  
email: daniel.himmelstein@gmail.com # suggested  
corresponding: true # optional, if set to true displays author's email for correspondence  
affiliations: # as a list, strongly suggested  
 - Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania  
 - Department of Biological & Medical Informatics, University of California, San Francisco  
funders:  
 - GBMF4552 # optional list of author's funding

Note that affiliations should be a list to allow for multiple affiliations per author.

### Thumbnail

A thumbnail is an image used to visually represent the manuscript, such as when a manuscript is shared on social media or added to the [Manubot catalog](https://manubot.org/catalog/). Specify a thumbnail in any of the following ways:

1. placing an image named thumbnail.png anywhere in the manuscript repository (for example, in the root directory).
2. setting thumbnail in metadata.yaml to a path, relative to the repository root, where the image file is located. Example: yaml thumbnail: build/assets/thumbnail-1000x1000.png
3. setting thumbnail in metadata.yaml to an absolute URL where the image is located. Example: yaml thumbnail: https://github.com/greenelab/meta-review/raw/master/thumbnail.png

Methods 2 and 3 take precedence over method 1. View the [guidelines here](https://github.com/manubot/catalog#thumbnail-guidelines) for suggestions on how to create a good thumbnail. Key points are that thumbnails should be 1000 × 1000 pixels, PNG formatted, and striking.

## Custom formatting

Modifying the manuscript formatting requires modifying the CSS in the file <build/themes/default.html>. Common formatting changes, such as [font size](https://github.com/manubot/rootstock/issues/239) and [double spacing](https://github.com/manubot/rootstock/issues/244), can be found by searching the [Rootstock issues](https://github.com/manubot/rootstock/issues). Open a [new issue](https://github.com/manubot/rootstock/issues/new) if you have a new formatting question.

Changing the citation style or which interactive HTML plugins are loaded requires editing the options specified by Pandoc defaults files in <build/pandoc/defaults>. The citation style is determined by the Citation Style Language file specified in [common.yaml](build/pandoc/defaults/common.yaml):

metadata:  
 csl: build/assets/style.csl

The value for metadata.csl can be a URL, allowing access to thousands of existing styles hosted by [Zotero](https://www.zotero.org/styles) or the [CSL GitHub](https://github.com/citation-style-language/styles). For example, the following options replace the Manubot citation style with the *PeerJ* style:

metadata:  
 csl: https://github.com/citation-style-language/styles/raw/906cd6d43d0c136190ecfbb12f6af0ca794e3c5b/peerj.csl

## Spellchecking

When the SPELLCHECK environment variable is true, the pandoc [spellcheck filter](https://github.com/pandoc/lua-filters/tree/master/spellcheck) is run. Potential spelling errors will be printed in the continuous integration log along with the files and line numbers in which they appeared. Words in build/assets/custom-dictionary.txt are ignored during spellchecking. Spellchecking is currently only supported for English language manuscripts.

## Manubot feedback

If you experience any issues with the Manubot or would like to contribute to its source code, please visit [manubot/manubot](https://github.com/manubot/manubot) or [manubot/rootstock](https://github.com/manubot/rootstock).

## Citing Manubot

To cite the Manubot project or for more information on its design and history, see @doi:10.1371/journal.pcbi.1007128:

**Open collaborative writing with Manubot** Daniel S. Himmelstein, Vincent Rubinetti, David R. Slochower, Dongbo Hu, Venkat S. Malladi, Casey S. Greene, Anthony Gitter *PLOS Computational Biology* (2019-06-24) <https://doi.org/c7np> DOI: [10.1371/journal.pcbi.1007128](https://doi.org/10.1371/journal.pcbi.1007128) · PMID: [31233491](https://www.ncbi.nlm.nih.gov/pubmed/31233491)

The Manubot version of this manuscript is available at <https://greenelab.github.io/meta-review/>.

## Acknowledgments

We would like to thank the contributors and funders whose support makes the Manubot project possible. Specifically, Manubot development has been financially supported by:

* the **Alfred P. Sloan Foundation** in [Grant G-2018-11163](https://sloan.org/grant-detail/8501) to [**@dhimmel**](https://github.com/dhimmel).
* the **Gordon & Betty Moore Foundation** ([**@DDD-Moore**](https://github.com/DDD-Moore)) in [Grant GBMF4552](https://www.moore.org/grant-detail?grantId=GBMF4552) to [**@cgreene**](https://github.com/cgreene).