**CORD19 - Data dictionary**

The COVID-19 Open Research Dataset (CORD-19) is a growing resource of scientific papers on COVID-19 and related historical coronavirus research. CORD-19 is designed to facilitate the development of text mining and information retrieval systems over its rich collection of metadata and structured full text papers.

CORD-19 is released weekly. Each version of the corpus is tagged with a date stamp.

The files in each version are:

* **changelog**: A text file summarizing changes between this and the previous version.
* **cord\_19\_embeddings.tar.gz**: A collection of precomputed [SPECTER](https://arxiv.org/abs/2004.07180) document embeddings for each CORD-19 paper
* **document\_parses.tar.gz**: A collection of JSON files that contain full text parses of a subset of CORD-19 papers
* **metadata.csv**: Metadata for all CORD-19 papers.

**Metadata.csv overview :**

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Data type** | **Description** | **Comments** |
| cord\_uid | str | Unique identifier to each CORD-19 paper | Not necessarily unique per row.As same paper can come from different sources. |
| sha | List[str] | Secured Hash Algorithm(SHA1) Id’s of all the pdf’s associated with CORD-19 dataset. | Mostly papers will have either zero or single value here .Some papers will have multiple indicating more than one pdf associated with single paper. |
| source\_x | List[str] | Sources of paper | Single paper can have multiple sources |
| title | str | Paper title |  |
| doi | str | Digital object identifier (DOI) of the paper |  |
| pmcid | str | Paper’s Id on PubMed Central | Begins with pmc followed by an integer |
| pubmed\_id | Int | Paper’s Id on PubMed |  |
| license | str | License associated with the paper |  |
| abstract | str | Paper’s abstract |  |
| publish\_time | Str | Published date of the paper | In yyyy-mm-dd  format |
| authors | List[str] | Authors of the paper | Each author name is in Last, First Middle format and semicolon-separated |
| Journal | Str | Paper journal | Strings are not normalized |
| mag\_id | Int | Value field for the paper as represented in Microsoft Academic graph | Deprecated |
| who\_covidence\_id | Str | ID assigned by the WHO for this paper |  |
| arxiv\_id | Str | arXiv ID of the paper |  |
| pdf\_json\_files | List[str] | Paths from the root of the current data dump version to the parses of the paper PDF’s in Json format | Multiple paths are semicolon-separated |
| pmc\_json\_files | List[str] | Same as above but corresponds to the full text XML files taken from PMC,parsed in to the same Json files as above | Multiple paths are semicolon-separated |
| url | List[str] | All URL’s associated with the paper | Semicolon-separated |
| s2\_id |  | Semantic Scholar ID for this paper | Can be used with the Semantic Scholar API (e.g. s2\_id= 1234 corresponds to http://api.semanticscholar.org/corpusid:1234) |