Dataset Descriptions

Use this document to provide descriptions of the entities and attributes of datasets in the IntelComp data lake.

Please also include information on the provenance of the dataset and how often it can be updated

[The metadata schemata can be found here.](https://docs.google.com/spreadsheets/d/19nx4mn0dfc--cMAIusK5L5UvjuQVgByh8KPZR28LM7g/edit?usp=sharing)

## UC3M - SemanticScholar

Provenance: [SemanticScholar](https://api.semanticscholar.org/corpus/download/)

Update frequency:

Source updates monthly

Proposed updates every year

We will use a coloring scheme for better identification of the most relevant fields

| Text Data  (For topic modeling, etc) | Time data | Regional Data | Other relevant fields (e.g., budget) | Unique identifiers |
| --- | --- | --- | --- | --- |

| **Entities**  # table entries: 211.633.022 | | | |
| --- | --- | --- | --- |
| **name** | **description / assumptions** | **format** | **foreign key / relation** |
| papers | Each of the publications. | parquet |  |

| **Attributes** | | | |
| --- | --- | --- | --- |
| **name** | **description / assumptions** | **format** | **foreign key/relation** |
| id | Unique S.S. identifier. | long |  |
| title |  | string |  |
| s2Url | URL to S2 research paper. | string |  |
| year | Publication year. | integer |  |
| doi |  | string |  |
| pmid | PubMed identifier. | string |  |
| magId | Microsoft Academic Graph identifier. | string |  |
| externalids | Other external identifiers available for the paper | struct:   * ACL: string * ArXiv: string * CorpusId: string * DBLP: string * DOI: string * MAG: string * PubMed: string * PubMedCentral: string |  |
| fieldsOfStudy | extracted from s2fieldsofstudy. Several models available, we keep s2-fos-model | array[string] |  |
| publicationtypes | JournalArticle: 40.043.054  Review: 14.740.426  Conference: 3.892.498  Study: 2.341.448  CaseReport: 2.183.428  LettersAndComments: 1.544.547  Editorial: 665.251  ClinicalTrial: 582.213  Book: 280.889  News: 239.700  MetaAnalysis: 103.700  Dataset: 727  (Many null, several + than one) | array[string] |  |
| publicationdate | Complete date of publication | string |  |
| journalName |  | string |  |
| journalPages |  | string |  |
| journalVolume |  | string |  |
| venue |  | string |  |
| publicationvenueid |  | string |  |
| isopenaccess |  | boolean |  |
| referencecount | Number of references included in the paper | int |  |
| citationcount | Number of times the paper has been cited | int |  |
| influentialcitationcount | Number of highly influential citations | int |  |
| paperAbstract | Available for 100.048.872 papers | string |  |
| openaccessinfo |  | struct:   * externalids: struct   + ACL   + ArXiv   + DOI   + MAG   + PubMedCentral * License: string * license: string * Url: string |  |

| **Entities**  # table entries: 81.067.677 | | | |
| --- | --- | --- | --- |
| **name** | **description / assumptions** | **format** | **foreign key / relation** |
| authors |  | parquet |  |

| **Attributes** | | | |
| --- | --- | --- | --- |
| **name** | **description / assumptions** | **format** | **foreign key/relation** |
| id | Unique identifier. | string |  |
| name | Author’s name. | string |  |
| aliases | Different names | array[string] |  |
| papercount |  | integer |  |
| citationcount |  | integer |  |
| hindex |  | integer |  |

| **Entities**  # table entries: 590.620.314 | | | |
| --- | --- | --- | --- |
| **name** | **description / assumptions** | **format** | **foreign key / relation** |
| paper\_author | Paper – Authors relationship | parquet |  |

| **Attributes** | | | |
| --- | --- | --- | --- |
| **name** | **description / assumptions** | **format** | **foreign key/relation** |
| paper\_id | S.S. paper identifier. | long | papers(id) |
| author\_id | S.S. author identifier. | string | authors(id) |

| **Entities**  # table entries: 2.557.754.271 | | | |
| --- | --- | --- | --- |
| **name** | **description / assumptions** | **format** | **foreign key / relation** |
| citations | Paper – Paper relationship | parquet |  |

| **Attributes** | | | |
| --- | --- | --- | --- |
| **name** | **description / assumptions** | **format** | **foreign key/relation** |
| source | S.S. paper identifier. | long | papers(id) |
| dest | S.S. paper identifier. | long | papers(id) |
| isinfluential | S2 model (only for full pdfs) | boolean |  |