Alyssa Schomber

University of Washington Information School

LIS 545 B Wi 23: Data Curation I: Fundamentals

Professor Matthew Mayernik

February 24th, 2023

**Report**

**Recommended Citation**

Dhruvil Dave. (2020). *New York Times Best Sellers* [data set]. The New York Times [producer]. Kaggle [distributor]. <https://doi.org/10.34740/KAGGLE/DS/1018720>

**Long-term Preservation**

The .csv file is one of the oldest formats and is likely to remain in use for quite some time; however, because it is not standardized, not all software or programs may be able to render it properly (or as it may have been intended). Instruction as to the programs or software that do display it correctly would help ensure future use, as it could be easier translated to different programs with minimal potential for corruption. A spreadsheet, database, or text editor software is required to open the dataset. However, Microsoft Excel, for example, does not display the data consistently, while Google Sheets does. An incorrect display of ISBN-13 identifier occurs in Microsoft Excel, while this does not occur in Google Sheets. Accessing this dataset in a text editor format is difficult to parse; however, it may be best to access the purest form of the data. Another issue is the amazon product links listed in the data. These are likely to terminate in the future and require a more secure URI.

**Copyright License**

This dataset had a previously assigned license, CC BY-NC-SA 4.0, which is still appropriate as set by the New York Times’ terms and conditions.

**Human Subject Considerations**

The only identifiable data in the set are the names of the works' creators, which are public knowledge. Due to this, no anonymization was required.

**Data and Metadata Profile**

**Data**

The New York Times Best Sellers data set consists of the top five books in each format and genre category listed on the "Best Sellers List published by The New York Times every Sunday" from January 3rd, 2010 to December 29th, 2019 (Dhruvil Dave, 2020). According to the metadata, the data originates from The New York Times Books API. The key stakeholders of this data include The New York Times & The New York Times Developer Network, The New York Times Company (the mass media publishing company), the data compiler owner (Dhruvil Dave), authors, publishers, booksellers, NYT readers, "the web developer community" and noncommercial users (FAQ | Dev Portal, n.d.).

There is one data file in a .CSV file format. In this file, there are twelve columns, which contain cells of text and string, integer, and datetime formulas.  The data is organized in the file by published\_date, list\_name, list\_name\_encoded, # rank, isbn13, isbn10, title, author, description, amazon\_product\_url, # price, and # weeks\_on\_list. The data set is contained by a .csv text file, which requires either a spreadsheet, database, or text editor software to open. However, upon opening the software in Microsoft Excel, some of the text displays incorrectly. Some dates found in the “published\_date” column appear as a string of hashes unless the column is widened, and the book identifiers provided in the “isbn13” column are transformed into E notation and cannot be obviously readjusted. However, when opened in Google Sheets, all text displays correctly without needing to modify it. This might insinuate that the data set is best accessed via this program.

The data license provided with the data set is the “CC BY-NC-SA 4.0" (Attribution-NonCommercial-ShareAlike 4.0 International),” which allows sharing and adapting, while restrictions require attribution, that the material be used noncommercially, that any changes to the material are redistributed under the same license, and that “you may not apply legal terms or technological measures that legally restrict others from doing anything the license permits” (Creative Commons, n.d.). The NYT Developer Network does not allow commercial use of data but allows noncommercial use (FAQ | Dev Portal, n.d.). Its “Terms of Use” spells out the general restrictions further, including: the previously stated prohibition of commercial use, not allowing use “in a manner that exceeds reasonable request volume, constitutes excessive or abusive usage, or otherwise fails to comply or is inconsistent with any part of the NYT API documentation,” or promotion, use or connecting of data to “spyware, adware, or other malicious programs or code, hate materials (e.g. Nazi memorabilia) or materials urging acts of terrorism or violence, hacking/surveillance/interception/ descrambling equipment, gambling items or non-transferable items such as airline tickets or event tickets” (Terms | Dev Portal, n.d.).

**Metadata**

The metadata is found on a dedicated Kaggle landing page for the data set itself, called a “notebook.” Kaggle allows the data compilers to provide information on collaborators, authors (& bio), coverage (temporal & geospatial), DOI citation (with a BibTeX & APA automatically generated citation), provenance (sources & collection methodology), license, and expected update frequency. There is also a content description and acknowledgements section at the top of the page. It has an activity overview, which shows data set, notebook, and discussion stats, as well as other top contributors to the data and charts tracking downloads and views. The creator of this data in particular supplies all the above except for geospatial coverage, author biography, and expected update frequency. However, in downloading the zipped file, none of this metadata is included.

Overall, the Kaggle metadata is very brief, but in general it does detail what the data is, whence it derives, who the creators are, what the rights are, and provides a DOI handle. According to Kaggle’s GitHub, which can be accessed through the public API section of their “How to Use Kaggle” page, Kaggle’s API uses a Frictionless Data data package specification for their metadata standard (Public API Documentation, n.d., Dataset Metadata, n.d., Frictionless Data, n.d.). Every single Kaggle data set entry uses this same data package specification on their notebook pages.

Although this data set has a DOI provided for it, unless one is looking for it on Kaggle, it is difficult to discover it. Greater linking or discovery tags in the metadata could better connect it to other repositories. Also, further descriptive metadata would be helpful for anyone unfamiliar with the data already. Notes explain the differences between hardback and paperback sales, ISBN-10 and ISBN-13 and why some cells lack them, as well as what program the data set would be best accessed through would be especially helpful.

**What publications have been written (if any) based on this dataset?**

There are no publications listed or provided with the data set. However, the working paper, “Politics and Children's Books: Evidence from School Library Collections,” referenced this data set, using it as one of many resources to “to identify patterns in library resources and content” (Mumma, 2022). This is discoverable by searching for the DOI provided in Google Scholar. Also, although not technically publication, Kaggle user JESSCPA23 evaluated, cleaned, and analyzed Dhruvil Dave’s data set (NY Times Bestseller List Analysis, n.d.). This can be found on the data set’s notebook, as any associated or related notebooks are documented in the “related notebooks” container beneath the activity overview.

**Bibliography**

*Creative Commons—Attribution-Noncommercial-Sharealike 4.0 International—CC BY-NC-SA 4.0*. (n.d.). Retrieved January 25, 2023, from <https://creativecommons.org/licenses/by-nc-sa/4.0/>

*Dataset Metadata*. (n.d.). GitHub. Retrieved January 27, 2023, from https://github.com/Kaggle/kaggle-api/wiki/Dataset-Metadata

Dhruvil Dave. (2020). *New York Times Best Sellers* [Data set]. Kaggle. <https://doi.org/10.34740/KAGGLE/DS/1018720>

*FAQ | Dev Portal* (n.d.). Retrieved January 25, 2023, from <https://developer.nytimes.com/faq#a2>

*Frictionless Data.* (n.d.).Frictionless Data. Retrieved January 27, 2023, from https://frictionlessdata.io/

*NY Times Bestseller List Analysis*. (n.d.). Retrieved January 27, 2023, from <https://kaggle.com/code/jesscpa23/ny-times-bestseller-list-analysis>

Mumma, K. S. (2022). Politics and Children's Books: Evidence from School Library Collections. Edworkingpapers.org.

*Public API Documentation*. (n.d.). Retrieved January 27, 2023, from https://www.kaggle.com/docs/api

*Terms | Dev Portal* (n.d.). Retrieved January 25, 2023, from <https://developer.nytimes.com/terms>

**Repository Profile**

I chose the data repository, Work With Data, which is “a data library, collecting data from multiple reliable sources and merging them altogether into one database. It focuses on access through insights, datasets and api” (Re3data.Org, 2022). It focuses on “empowering everyone with open data combined from diverse sources & augmented with AI” (Work with data, n.d.).

**Why did you choose this repository?**

I chose this repository because my dataset is humanities/literary-related, and it was difficult to find a repository that was both broad and specific enough to allow for it. In the end, I chose Work With Data because it—much like Kaggle, the repository I obtained it from—was a general, catchall repository, and according to re3data.org, deals with subjects such as "humanities,” “humanities and social sciences," and “literary studies” (2022). Fundamentally, Work With Data is thematically interoperable with Kaggle, which “supports a variety of dataset publication formats,” includes a great variety of datasets, and accepts similar file formats (Datasets documentation, n.d.). For example, similar datasets already uploaded to Work With Data include a dataset on the works of Agatha Christie, and many of which were taken from a literary studies-focused source like the British Library (Books where author includes agatha christie, n.d., The British Library, n.d.).

**Is the repository open for data submissions from anybody, or does it have a defined collection scope?**

Anyone can submit data so long as they register an account. There are also pay-to-use services in the form of credits, which “unlock the use of AI generated data,” offered in an individual or enterprise plan (Credits, n.d.). Although this credit system is presented as an add-on service and is not necessary to upload or download data, its very existence may limit the user and submission base depending on individual financial circumstances.

**What data will the repository accept, e.g. are there any stated limits to what can be deposited? Comment on any limits related to data types (e.g. observational, simulation, etc), domain (biology, social science, etc), or file formats.**

Any data type or domain seems to be allowed. However, .CSV and .XLSX are the only file formats accepted for datasets, while .PNGs are the only file formats accepted for images or graphs.

**What guidance does the repository provide to a potential data submitter as to what should be in the “Submission Information Package” (SIP)\*?**

There is no public guidance beyond the aforementioned file format limitations.

**Does the repository provide any human assistance or consulting to the submitter?**

The repository does not have any direct assistance or consulting for the submitter. There is only a contact email: hello@workwithdata.com.

**Does the repository require metadata to be submitted in any specific structure, or according to any specific standard?**

No, it seems like the metadata is appended via "tags" by the user on the dataset’s landing page.

I**s a login required to download data? If so, what is the process for creating a login?**

There is a login required to download data. You can sign up for free by entering your email and creating a username and a password (Sign up for free, n.d.).

**Is more than one access mechanism provided? E.g. direct file download, database query, automated script, etc)**

Direct download is the only access mechanism provided.

**Does the repository display metadata using any specific metadata standard?**

There appears to be no metadata standard in use. Each dataset has its own assigned metadata by its creators. Dashboards seem to be generated automatically and compile/track certain data points as they appear in the repository. Each dashboard uses its own specific metadata logic and adapts to the subject, i.e. "Jane Austen" includes metadata on editions, date range, and so on, while "Jacksonville" includes metadata on population, companies, and so on (Jane Austen, n.d., Jacksonville, n.d.). Also, according to the icons on re3data.org’s Work With Data entry, "the research data repository is neither certified nor supports a repository standard" (2022).

**What is included in this repository’s “Dissemination Information Package” (DIP)\*?**

By observing the entries, the DIP consists of a landing page with the title of the dataset, the number of rows it consists of, the full dataset or sample dataset in a .CSV file, a dataset preview consisting of 15 rows, and a column that includes data editing capabilities. For images, a description is included, as well as metadata containing dates, measurements, details, provenance, location, and history, as well as "context" (which is only accessible with an account).

**Bibliography**

*Books where author includes agatha christie.* (n.d.). Work With Data. Retrieved February 10, 2023, from https://www.workwithdata.com/dataset

*Credits*. (n.d.).Work With Data. Retrieved February 10, 2023, from https://www.workwithdata.com/credits

*Datasets documentation.* (n.d.).Retrieved February 10, 2023, from <https://www.kaggle.com/docs/datasets>

*Jacksonville*. (n.d.). Work With Data. Retrieved February 10, 2023, from https://www.workwithdata.com/city/46275

*Jane Austen.* (n.d.). Work With Data. Retrieved February 10, 2023, from https://www.workwithdata.com/author/jane-austen-1775

Re3data.Org. (2022). *Work with data*. 5.47 million entities; 4 billion data points. <https://doi.org/10.17616/R31NJN7H>

*Sign up for free*. (n.d.). Work With Data. Retrieved February 10, 2023, from https://www.workwithdata.com/signup

*The British Library*. (n.d.). Work With Data. Retrieved February 10, 2023, from <https://www.workwithdata.com/source/british-library>

*Work with data*. (n.d.). Work With Data. Retrieved February 10, 2023, from <https://www.workwithdata.com/>