Data and Metadata Profile

The selected data set is a collection of all of streaming service Netflix's available television shows and movies, dated to 2020 and hosted on Kaggle. This data includes the many facets of these media items, such as their title, year of release, country of origin, type of media (show or movie), duration, rating, and description. On the streaming side, it displays the date Netflix added the media (*Netflix Movies and TV Shows*, n.d.). This data set has enormous potential, through both itself and through comparisons with similar sets, to demonstrate the trajectory of the popular streaming service's offerings. These potential trends include TV shows overtaking movies, quality of releases per IMDB or Rotten Tomatoes, availability of older intellectual properties, competition shrinking the overall library size, etc.

This data set, compiled by Shivam Bansal, uses the website Flixable.com as a source. Flixable.com creates up-to-date lists of all of the media available on Netflix, including such metadata as the official description, date added, release date, director, cast, genre, and country of origin. Other features include on-site ratings, IMDB ratings, and an ability to assemble a list of media to watch (with the creation of an account) (*Full List of Movies and TV Shows on Netflix | Flixable*, n.d.). Both Flixable and Bansal are stakeholders, as Flixable provides the data that Bansal compiles into the data set. Of course, there are secondary stakeholders in this data set, such as media companies, competing streaming services (e.g., Disney+ and Amazon Prime), and consumers looking at their options for programing. Netflix themselves might also be a secondary stakeholder, as Flixable denotes their holdings and can either persuade or dissuade customers from buying or renewing a subscription. Such a data set can have significant ramifications for both observers and for the providers themselves.

The data set hosted on Kaggle has only one file, a .csv downloadable that is both viewable on Kaggle and in desktop form. This file contains only a single spreadsheet, with unique identifiers for each piece of media on Netflix. The unique identifiers sort the spreadsheet in alphabetical order, but it is easy to change the sorting order on both Kaggle and through spreadsheet applications, such as Microsoft Excel or Google Sheets. It is helpful that Bansal limited himself to just one file, as it makes searching and sorting the 7,787 entries easier/customizable to the researcher's preference.

The data set has no rights restrictions. In the metadata, Bansal lists the licensing as CC0: Public Domain. According to the Creative Commons website, this means that the creator has waived the copyright on the item (*Creative Commons — CC0 1.0 Universal*, n.d.; *Netflix Movies and TV Shows | Kaggle*, n.d.). Anybody can use the data Bansal compiled from Flixable, which gives it plenty of flexibility for future publications. While Bansal did not present the data with any publications, the set should prove useful to publications.

The data, despite being easy to sort and very descriptive, has some notable flaws that, if corrected, would improve the set. Flixable lists IMDB ratings within their listings, which would give an indication of the quality of the media in the set. This should be an easy fix to add this information, perhaps doable with a crawl script building on top of the existing spreadsheet. Outside of Flixable's given information are elements like Rotten Tomatoes scores, numbers of episodes for TV shows, and production company. Such information would also be relatively easy to add, likely through importing a data set sourced from sites like Rotten Tomatoes or IMDB.

With regards to metadata, the information is a listing on the Kaggle site. It contains the license, visibility, source, collection method (regular API calls), creator, expected update frequency (monthly), last updated, date created, and current version (for this set, Version 4 is the

latest). On the “data” tab on the Kaggle site, the data set’s column metadata appears. This metadata features the average attributes of each column (*Netflix Movies and TV Shows | Kaggle*, n.d.). This metadata does not seem to fit any predetermined schema, save for the standard requirement of data sets on Kaggle (usage information, provenance, maintainers, and updates). It is not particularly exhaustive, but it provides enough of a reference for researchers to confidently use its information and know that it is relatively up-to-date, free to use, and has a reliable source for data. Adding more release notes or expanding upon the column averages would be a welcome implementation, as would a rich text file denoting version changes/improvements and averages of data (for offline use, without Kaggle’s averages).

As far as representation in publications goes, this dataset was featured in both an article for *Medium* and in an SQL coding language project. The former, “A data-driven look at what’s worth watching on Netflix” by Sriram Sharma, details the data set’s usefulness in illustrating the quality of content on Netflix. Combining it with an IMDB rating data set, he determined that Netflix does not feature many movies or TV shows rated above 8.0 on IMDB (Sharma, 2020). The SQL project demonstrated how the dataset’s well-produced nature made it easy to use for searches in SQL editors/viewers (*SQL Project 2) Netflix Movies and TV Shows*, n.d.).

Despite these usages, the data set has not been featured in any scholarly publications, as determined by my searches on Google Scholar. Despite its relative obscurity, it should be useful to professional researchers trying to analyze trends within Netflix’s programing or how streaming has exposed previously-existing intellectual properties (i.e., not original programing) to new audiences. It could even be used as a key source in determining online engagement with specific intellectual properties, by corroborating the spreadsheet with IMDB ratings (as Sharma did), social media traffic for the topic (e.g., Twitter and Reddit posts), and Nielsen ratings (for

pre-existing IP's). The possibilities with Bansal's data are enormous and, with Bansal continuing to update the data, can extend beyond a limited time scope.

Repository Profile

The repository selected for the chosen dataset is the Data and Service Center for the Humanities (DaSCH). Based in Switzerland, the DaSCH's small set of collections center around projects involving the humanities and cultural preservation. Additionally, the DaSCH also prides itself on preserving of data and making it accessible for additional researchers. The selection came about after a long search for a humanities-related repository that would allow databases and had a wide enough breadth to include more obscure/technical data like the Netflix listings dataset (*Data and Service Center for the Humanities (DaSCH)*, n.d.). Since there are no specific streaming/technology platforms listed on the Registry of Research Data Repositories, it made sense to import the dataset into a more general research-focused institution.

This repository skews towards humanities and cultural heritage data, the specific nature of which is variable. Examples of these collections include HyperHamlet, "a hypertext of Hamlet, it gives access to thousands of extracts from later texts that quote particular lines," and the Photocollection of the Swiss Society of Folklore (*Projects Hosted by DaSCH*, 2019). There is a push towards contextualizing media and using it to tell histories of specific places and texts. Hosting the Netflix listings here would be unorthodox for their collections (their primary focus is on European cultural heritage). Still, it fits within the guidelines of submissions. It can also provide its own "history of media," as represented by the aggregate listings of a popular streaming platform in the United States.

DaSCH primarily deals with qualitative research data relating to humanities projects, with bitstream object linking to multimedia. The repository allows multiple functions, including the submission of "simple" datasets in the form of flat files (e.g., a .csv database, like the Netflix listings) alongside complex datasets that include additional multimedia materials. However, the DaSCH will remodel data to be compliant with its platform for ease of curation and adaptation to new technological developments and the DaSCH standard. Since the repository maintains the data to the OAIS standard, the original data will remain for researchers to retrieve if possible. On top of that, the repository provides archival resource keys (ARKs) as permanent ID's for the data and any digital objects included. (*DaSCH Services*, 2016). Therefore, not that much data needs adapting to the repository.

The repository primarily provides support through either collaboration with an institution's research and IT departments (specifically in Switzerland, where such extensive support will be easier for the organization) (*DaSCH Mission Statement*, 2016). However, if there is no institution or a research/IT unit, DaSCH can provide direct support for scholars. Their services include consulting on data management/creation of digital objects (e.g., digitization and workflows for collection) and data modeling advice. DaSCH will also develop tools for preparing data for easy dissemination. The repository will also offer training for institutions to use the repository and its service platform (*DaSCH Services*, 2016). Due to this extensive support, using this repository for the Netflix listings would be advantageous, as the direct assistance might extend across borders and optimize the dataset for hosting.

When considering the metadata standards, the DaSCH does not list any specific metadata as the repository standard. However, re3data.org claims that the repository uses Dublin Core

(DCC), which is a well-recognized database schema and would be easy to implement for the dataset (*Data and Service Center for the Humanities*, n.d.).

Data access is a fair bit more complicated for some projects. While many of the projects (such as HyperHamlet) have their own sites to present the data and interact with it, there are no ways listed to download information or database files on the public site. All items are accessible through database queries using built-in search engines. This absence is somewhat understandable, given the potential licensing issues with digital objects (as not all items are in the public domain/creative commons licensed).

The service platform has a place where a user could log in, but no way to register an account visible on the public interface (*Bilddatenbank Bibliothek St. Moritz*, n.d.). Therefore, one would presumably need to contact the repository to get a login and have a dataset ready to contribute. The front page of the website does state that prospective data project owners should email the repository. Owners must provide both a description of the project and a Data Management Plan (DMP) to demonstrate an owner's commitment to the data/frequency of updates and distribution methods (*Data and Service Center for the Humanities (DaSCH)*, n.d.). After this, one would presumably receive an account as an owner. Downloading entire databases could occur then, but there is no explicit information about it or what form it would take (e.g., automated scripts or direct file downloads).

When considering the dissemination information package, the discoverable information has both digital objects and their respective metadata. The DaSCH also provides both ARKs and access points for all material for both ease of discoverability and accessibility, and interacting with other collections of DaSCH. It also uses linked open data technology like RDF to increase interoperability and interactions with other platforms (*DaSCH Services*, 2016). This information

available on the public interface provides comprehensive information, which should satisfy researchers. Granted, having .csv or similar database files readily available would be the best option. Still, the database files could appear on the collection's specific site in addition to the search engine section through the public interface design.

The DaSCH should be the best possible host for the Netflix listing dataset, given its open nature and the support it can provide for shaping the dataset to the most optimal standards for discoverability and interoperability. While there may be some adaptations needed and a well-structured argument for the dataset to join the repository, the DaSCH will be ideal.

Recommended Data Citation

Shivam Bansal (2020). Netflix Movies and TV Shows (Version 4) [Dataset]. Retrieved from https://www.kaggle.com/shivamb/netflix-shows

Long-term Preservation Concerns

Due to the continuously-updated nature of the dataset, the owners and contributors of the dataset will update the files upon the release of each new version of Bansal's dataset. Not much work has to be done, as updates only involve uploading the new .csv file (and .xlsx derivative) and adjusting the metadata and report files to reflect new information. Since the dataset uses .csv, .txt, .xlsx, and .pdf files, there is limited danger of any of the file formats becoming obsolete or incompatibility with modern computer systems.

Copyright

Since the original dataset is public domain, this github-hosted version of the database will also be public domain.

Human Subject Considerations

Due to the nature of the data, no human subject considerations are needed for the dataset. If Netflix or Flixable requests the takedown of the dataset, the owner and contributors will argue the public domain nature of the list of names and publicly-available information.

References

Bilddatenbank Bibliothek St. Moritz. (n.d.). Retrieved February 11, 2021, from https://data.dasch.swiss/

Creative Commons—CC0 1.0 Universal. (n.d.). Retrieved January 28, 2021, from https://creativecommons.org/publicdomain/zero/1.0/

DaSCH Mission Statement. (2016, December 23). Data and Service Center for the Humanities (DaSCH). https://dasch.swiss/mission/

DaSCH Services. (2016, December 30). Data and Service Center for the Humanities (DaSCH). https://dasch.swiss/services/

Data and Service Center for the Humanities. (n.d.). Re3data.Org. https://www.re3data.org/repository/r3d100012374

Data and Service Center for the Humanities (DaSCH). (n.d.). Data and Service Center for the Humanities (DaSCH). Retrieved February 11, 2021, from https://dasch.swiss/

Full List of Movies and TV Shows on Netflix | Flixable. (n.d.). Retrieved January 28, 2021, from https://flixable.com/

Netflix Movies and TV Shows. (n.d.). Retrieved January 28, 2021, from https://kaggle.com/shivamb/netflix-shows

Netflix Movies and TV Shows | Kaggle. (n.d.). Retrieved January 28, 2021, from https://www.kaggle.com/shivamb/netflix-shows/metadata

Projects hosted by DaSCH. (2019, September 29). Data and Service Center for the Humanities (DaSCH). https://dasch.swiss/projects/

Sharma, S. (2020, February 18). *A data-driven look at what's worth watching on Netflix*. Medium. https://svsharma.medium.com/a-data-driven-look-at-whats-worth-watching-on-netflix-7f5e13db027e

SQL Project 2) Netflix Movies and TV Shows. (n.d.). 네이버 블로그 | Coding Studying Log. Retrieved January 28, 2021, from https://blog.naver.com/eommelissa9813/221891112140