Data and Metadata Profile

Through FigShare, I accessed the Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach dataset (https://figshare.com/articles/dataset/_Personality_Gender_and_Age_in_the_Language_of_Social_Media_The_Open_Vocabulary_Approach_/808535). In this dataset, 700 million words, phrases, and topic instances were taken from the Facebook messages of 75,000 volunteers. These volunteers then took personality tests. Using an open-vocabulary developed by the team behind the study, the language used by participants was then tied to their personality test results and demographics. The data comes from a group affiliated with the Positive Psychology Center at the University Pennsylvania. There are a number of stakeholders who have a vested interest in this dataset, including the 75,000 volunteers, the research group, the Positive Psychology Center, The University of Pennsylvania, and the Computer and Information Science department in The University of Pennsylvania who was also involved in this work.

This data set contains four files. The first is a tagged image file (tif) of a line graph correlating the sample size of volunteers and the significant correlated features of speech noted by the group along the lines of that sample size's age, gender, and Big Five personality test results. The second is a tif of word clusters representing words and phrases that distinguish groups with high agreeableness, low agreeableness, high conscientiousness, low conscientiousness, high openness, and low openness. The third file is a Microsoft Excel worksheet showing the twenty most prevalent words and their frequencies for each of the topics the group identified through Latent Dirichlet Allocation. The last file is a pdf of another table showing the prediction results for determining personality features based on language analysis. These files can be opened using a Windows operating system with the Windows Photo Viewer, Microsoft Excel, and an internet browser or Adobe Acrobat. The data is licensed by Creative

Commons by 4.0 Deed. This means that users are able to share and build upon the data for any purpose as long as they give credit and do not add additional restrictions on the data (CC by 4.0 Deed).

There is not much metadata for the files shared on figshare. Each of the files has the file name, size, and MD5 checksum number noted on figshare. Figshare shares the date and time that the dataset was shared along with the authors. The dataset is categorized with 'Biological Sciences' and 'Science Policy' on figshare. The only keyword attached to the dataset is 'open-vocabulary. More metadata can be found in the files themselves. The tif files record the title, copyright, dimensions, resolution, bit depth, compression, and resolution unit of each image. The pdf includes the name, type, and attributes. Lastly, the excel file records the username of the last person who saved the file, the revision number, the date and time last saved, the content type, the name, the type, and the attributes. Although the files have more metadata records than are available on figshare, there is still significant amounts of information missing. This is especially true for the tif files and the pdf, as there is no information about when they were originally created or by whom. The metadata which is recorded does not seem to be structured according to any metadata standard.

The lack of metadata included in the dataset leaves a lot of room for enrichment. In order to improve the discoverability of the dataset in the repository environment, more information should be included on figshare. This dataset is categorized under 'Biological Sciences' and 'Science Policy' but these are not related to the actual contents of the dataset. The dataset should be included under more accurate categories such as 'Human Society,' 'Language, communication and culture,' and 'Psychology.' More tags related to the dataset can be included for discoverability as well, such as 'Facebook,' 'social media,' 'linguistics,' 'personality,' etc.

When the dataset is being accessed for new purposes, descriptive titles and heading names are important. Looking at the data itself, it is difficult to tell what the figures and numbers mean. Headings such as ‘w1’ and ‘category’ are used without any documentation explaining what the numbers under those headings relate to. In order to understand the contents of the dataset, the user would need to be familiar with the associated article.

The article, “Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach,” is linked in the dataset. Beyond this, there are no other citations or publications associated with the dataset. Figshare shows that the dataset has been viewed 2,863 times and downloaded 377 times. However, when I searched for the article again on Google, I came across the same article on PubMed. This showed 173 articles had cited the data. The article associated with the dataset is being cited but the actual dataset on figshare is not. Figshare tells us that citations are measured using a research information database called Dimensions. It is not clear how the citations are measured and if it is possible that there are citations not measured by figshare (Usage Metrics).

Repository Profile

The dataset “Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach” is associated with a study in which the Facebook messages of 75,000 volunteers were analyzed using an open-vocabulary approach and tied to their gender, age, and personality test results (Schwartz). This dataset can be associated with several academic disciplines including linguistics, sociology, and psychology. Although the group who created the study is based at the University of Pennsylvania, the volunteers used in the study are not from any specified region. Based on these considerations, I wanted to choose a disciplinary repository without any regional associations. I have chosen to store the database in the repository

“PsychArchives.” PsychArchives is not tied to any specific region and is a “disciplinary repository for psychological science and neighboring disciplines” (PsychArchives).

In order to successfully submit data to PsychArchives, it must meet a number of requirements. The data must be related to psychology or a similar discipline. It must belong to one of the 20 digital research object types that the repository accepts. If you are submitting research data, it must include a codebook and the dataset must be anonymized. The data cannot be more than 1GB. The data just be in either English or German. It should include metadata which aligns with PsychArchives’ metadata schema. Finally, the data should be in file formats for long-term archiving.

PsychArchives accepts 20 different kinds of digital research objects. The scope of these objects is quite broad, including books, articles, code, images, sound, tests, research data, ‘other,’ etc. The chosen dataset would fall under the ‘research data’ object type. When submitting the data, PsychArchives walks you through the process using the ‘PsychArchives Submission Assistant,’ which requires an ORCID or Leibniz Psychology account. Before entering the Assistant page, PsychArchives also links potential submitters to their guidelines and informational pages such as the sharing level page, the quality and technical guidelines page, the DRO type overview page, etc. In terms of assistance, it seems that only the PschArchives Submission Assistant is available. There is an e-mail provided but this is only for those attempting to publish data larger than 1 GB. After the data is submitted, it will be reviewed by the PsychArchives Team (Contribute).

If you are not submitting data, but would like to use what’s available on the repository, a login is not required. Instead, a user clicks ‘download’ on the data they would like access to and is taken to the “Download Request” page. This page states that “The person who provided this

digital research object (contributor) wants to learn about the intended use of the requested file" (Download Request). You are able to write the reason why you want to use the data or you can click a box that says "I do not want to specify the intended use" (Download Request). After you write your message or click this box, you can click the button on the bottom of the page and your download will begin immediately.

Directly downloading the data in this way is one way in which users can access materials on PsychArchives. However, if you do not wish to download the data users may also click the preview button. The preview button allows users to view the data in the browser as if it were a pdf. But the user will not be able to download this view of the data. Other than a direct download or this preview feature, there are no other access mechanisms available on this data repository.

Even before downloading files from the repository, you are able to access the metadata of an object. PsychArchives displays metadata using their own metadata scheme. According to the metadata schema, each contribution must display the Author(s)/Creator(s), the Abstract or Description, a Persistent Identifier such as DOI, the Date of First Publication, the Publisher, and a Citation. The metadata of an object will also display the PsychArchives Acquisition timestamp, when it was made available on the repository, the object's publication status, the language of the content, Dewey Decimal classification numbers, a title, the DRO type of the file, and Leibinz Institute names or abbreviations (Baier).

Searching for the term "Dissemination Information Package" does not yield any results on PsychArchives. However, there is information on how data on the repository is to be disseminated. Under each download, there are two links. One is the sharing level of the object and the other is the licensing.

The sharing level link shows the sharing level of the object you are viewing but it is hyperlinked to the sharing level page on PsychArchives to provide more information about what these sharing levels mean. On PsychArchives objects can fall under public use or scientific use. If an object falls under scientific use, it means that it can only be used for scientific analysis or discourse (Sharing Levels). PsychArchives says that there are more sharing levels being planned but they are not yet available on the website.

The licensing link under an object will bring you to a page for the specific license placed on that object. There are four licenses which PsychArchives supports: Share Alike/Copyleft, Attribution, No rights reserved, and Scientific Use. The first three of these licenses are connected to a different website which describes the license (Rights and licenses). However, the Scientific Use license is created by PsychArchives itself. It is treated as its own data object by the repository and can be downloaded and cited like other objects (Kreutzer).

Additional Information

When citing the Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach dataset, there are two recommendations to keep in mind. The first is that if you are using the 'Cite' button on Figshare, it will direct viewers to the article associated with the dataset but not the dataset itself. It is recommended that if you are citing the dataset you should include the URL of the dataset on Figshare instead of the DOI that will redirect viewers to the article. However, if the dataset is available through PsychArchives it will be assigned a DOI (About PsychArchives). In this case, the DOI should be included in the citation.

The dataset includes three different types of files: two .tif files, a Microsoft Excel worksheet, and a .pdf file. The .tif file format is developed and supported by Adobe (TIFF Files) while Microsoft Excel worksheets are supported by Microsoft. .tif files and .pdf files can be

opened by many different types of software but a Microsoft Excel worksheet requires Excel. These factors all pose challenges to long-term preservation as the files might no longer be usable if the developing company is dissolved or if the required software to open the files is no longer available.

When licensing the dataset, CC by 4.0 is what is used on Figshare. This license is also supported by the PsychArchives repository for a public use sharing level (Rights and Licenses). CC by 4.0 is an open access license which I would recommend for this dataset as it allows the data to be used by the public and is compatible with both Figshare and PsychArchives.

In this dataset, information from the Facebook messages of 75,000 volunteers is used. In the paper associated with the dataset, there is no description of how these human subjects were protected other than to say that the study was approved by the University of Pennsylvania IRB and that the volunteers provided written consent (Schwartz). However, in the dataset there is no information about these volunteers provided. There are only counts of terms that they used. In two of the figures, these terms are associated with the volunteers' Big Five personality test results, but there is no information about how many volunteers got what results.

Works Cited:

About PsychArchives. Leibniz-Institut für Psychologie. https://www.psycharchives.org/en/about. 22 Feb. 2024.

Baier, C. et al. *PsychArchives Metadata Schema*. Leibniz-Institut für Psychologie, 10 Jan. 2024. https://doi.org/10.23668/psycharchives.14061. 9 Feb. 2024.

"CC by 4.0 Deed." Creative Commons, https://creativecommons.org/licenses/by/4.0/. 26 Jan. 2024.

Contribute. Leibniz-Institut für Psychologie. https://www.psycharchives.org/en/contribute. 9 Feb. 2024.

Download Request. Leibniz-Institut für Psychologie. https://pada.psycharchives.org/ac_zero_plus/15025ec6-eba3-453c-9611-1031d1e3abef. 9 Feb. 2024.

Kreutzer, T. *License for scientific purposes ("Scientific Use License")*. PsychArchives, 1 Jul. 2021. https://doi.org/10.23668/psycharchives.4988. 9 Feb. 2024.

PsychArchives. Leibniz-Institut für Psychologie. https://www.psycharchives.org/. 9 Feb. 2024.

Rights and Licenses. Leibniz-Institut für Psychologie. https://www.psycharchives.org/en/about#rights. 22 Feb. 2024

Schwartz H., et al. *Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach.* Figshare, 25 Sep. 2013. https://figshare.com/articles/dataset/_Personality_Gender_and_Age_in_the_Language_of_Social_Media_The_Open_Vocabulary_Approach_/808535. 9 Feb. 2024.

Schwartz H., et al. *Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach.* PLOS ONE, 25 Sep. 2013. https://doi.org/10.1371/journal.pone.0073791. 22 Feb. 2024.

Sharing Levels. Leibniz-Institut für Psychologie. https://www.psycharchives.org/en/about#sharing_levels. 9 Feb. 2024.

Tiff files. Adobe. https://www.adobe.com/id_en/creativecloud/file-types/image/raster/tiff-file.html. 22 Feb. 2024.

"Usage Metrics." Usage Metrics, Figshare, 2024, https://help.figshare.com/article/usage-metrics. 26 Jan. 2024.