Data and Metadata Profile

Selected Indicators from World Bank 2000-2021

Roberto Lofaro, Kaggle

https://www.kaggle.com/datasets/robertolofaro/selected-indicators-from-world-bank-20002019/data

These data are a collection of 34 "Indicator" statistics for the countries in the world. That data was originally from the World Bank, then were selected and compiled into a dataset by Robert Lofaro on Kaggle. The data contain statistics about important information such as inflation of consumer prices, labor force participation rate by gender, and strength of legal rights. This makes it valuable for potential key stakeholders including the World Bank, other international organizations, national governments and policy makers, journalists, and researchers.

This dataset contains 3 files – facttable.csv with the main data of the indicators, dimension_indicator.csv explaining indicator abbreviations and meanings, and dimension_country.csv explaining country abbreviations. The files are in CSV, or Comma-Separated Values format, a common text format supported by most spreadsheet and database management systems such as Microsoft Excel and Notepad. For my own viewing I used Excel, which formatted the data into organized tables for viewability.

The data was posted under a Attribution 4.0 International (CC By 4.0) license, meaning that it can be shared and adapted by other parties as long as proper credit is attributed.

The metadata listed on the Kaggle repository itself are quite limited, containing only the author name, dates of coverage, source, and methodology. The two "dimension" files in the database provide important metadata in explaining what the abbreviated code means in the main dataset, making it understandable to other users. These metadata are crucial for comprehending and potentially reusing the data, but are of limited use for other purposes such as evaluating the source quality of the data or details about its meaning. More comprehensive metadata is only found through the World Bank's online index. Lofaro's dataset contains the "keycodes' needed to be inputted into the World Bank catalog in order to discover the "raw" metadata (though the link to a specialized search from the Kaggle posting does not seem to work, and a manual search of the World Bank catalog is required instead). These metadata contain more in-depth information about the data collected, such as the original source, date collected, and other relevant information about the countries covered by the data. The metadata does not appear to be structured according to any metadata standard, containing a variety of idiosyncratic categories, not all of which are completely recorded for each country.

It is unclear how exactly Kaggle's repository search functions. It would be easier to discover this dataset if each of the 34 indicators were tagged; I initially discovered it through a simple keyword search for "legal" within Kaggle, matching one of the indicators and not matching the title, description, or listed metadata for the database, meaning that it is possible the indicators are already tagged for discovery purposes.

It is a somewhat elaborate process to find the "raw" metadata files, requiring searching the World Bank catalog with the appropriate abbreviation keycodes. Otherwise, it is fairly clear, if a little cumbersome, to figure out how to use the base file on a basic level using the two "dimension" files explaining the data presented within. It depends on what new use a user may have for the data which process they would require.

The uploader of the dataset, Roberto Lofaro, has indicated in the description that he intends to use it for a long term publishing project, which appears to not have been released yet at this point. He also uses it for regular article posts on his website, linked in the database description. Other than that, a general web search indicates no other citation of this particular dataset, though variations of the original World Bank data feature in many news and academic publications.

Repository Profile

For the dataset “Selected Indicators from World Bank 2000-2021”, I chose the Data Sharing for Demographic Research (DSDR) repository, part of the University of Michigan Institute for Social Research, and housed within the Inter-university Consortium for Political and Social Research (ICPSR).

https://www.re3data.org/repository/r3d100010256

https://www.icpsr.umich.edu/web/pages/DSDR/index.html

I selected this repository because it seemed fitting for the content of my chosen dataset, focused on demographic and social research statistics from around the world. It seemed to be the most broad and open out of the repositories on the topic listed in within the Registry of Research Data Repositories, unlike others which were narrowly tailored to a specific project, geographical region, or institution.

The repository encourages submissions of data from anyone, though when applying to submit there is a distinction made between logging in with an academic institution ID as opposed to a personal email. The DSDR has an Acquisitions Policy describing the type of data accepted – that which can be generalized to populations, and is primarily focused on population dynamics with a particular focus on public health. Priority is given to data from sources which, like the DSDR, is tied to the NICHD Population Dynamics Branch. Though my selected dataset would not have NICHD priority, and is not exclusively focused on public health, I believe that there are enough data within that could be potentially used for large-scale public health research, such as global population, wealth, and development statistics.

The DSDR encourages that data is submitted as SAS, SPSS, or Stat files, though ASCII with data definition statements is also acceptable. The repository indicates that datasets in other formats could be accepted as well, but does not elaborate on what criteria it would use for making a decision about other formats. For all formats, the DSDR requires that variable labels and value labels should provide clear description, missing data codes should be defined, and when applicable that data should be scrubbed of identifying information to maintain confidentiality.

The repository also expects metadata in the form of documentation files explaining the data, such as codebooks, data collection instruments, bibliographic information about the data, project summaries, or other information. This documentation should be submitted as Microsoft Word, ASCII, or DDI XML files, though again a vague allowance for other formats is potentially possible. Also required is a study description with summary information about the data collection following the ICPSR metadata schema format (https://icpsr.github.io/metadata/icpsr_study_schema/).

The DSDR heavily promotes the human assistance it offers in managing and enhancing the data, such as fixing it to be more readable and consistent, making it more discoverable, and assisting with disclosure issues. Staff is also available to contact for questions about other issues.

Downloading data requires agreement to a Terms of Use about privacy of research subjects, redistribution of data, citations, and other issues, as well as a login. Creating a login requires providing either a personal email or a centralized sign-in through an institution (though I received an error when attempting to register through my University of Washington account). Multiple methods of download are supported, including SAS, SPSS, Stata, R, ASCII, Delimited, and an option for documentation only. An option for a SDA analysis online is also available, but does not seem to be supported by most of the datasets in the repository. Metadata is available for download in four standards – Dublin Core, DDI 2.5, DATS 2.2, and DCAT-US 1.1. On each collection page is also a "Variables" tab explaining each variable label and its place within the datasets, which also qualifies as a form of metadata.

I do not know for sure whether my chosen dataset would actually be accepted by the DSDR repository. I believe that it fits well, but it is not from a priority source of the repository and there is some ambiguity over whether the subject is closely related enough to the DSDR's primary mission (though other data collections with similar focus do seem to be accepted). There are also potentially some issues over rights and citation, as the data of this set was originally researched by the World Bank and then worked on by a Kaggle user, rather than original data I had collected myself. However, this repository still ultimately seems to be the best fit with its subject matter, extensive assistance options, and wide range of supported formats.

Additional Information

1. Recommended Data Citation: For this dataset, I would use a basic DataCite format –

Lofaro, R. (2022). *Selected Indicators from World Bank 2000-2021*[Data set].

There is no DOI indicator for this dataset but perhaps one could be applied for if it were submitted to a repository such as the ICPSR.

2. Long Term Preservation: For the most part this repository is suited for long-term preservation. The file types included – text, XML, CSV, and PDF – are all viewable through easily available

software. Anyone wishing to access the repository through Github should be able to view the data within as long as Github itself is still operational. The most prominent issue with preservation is not within the repository itself, but with finding the raw metadata of the research statistics from the World Bank. This depends upon the World Bank's website continuing to be accessible. There already seem to be some differences in the site's format when I operated it compared to what was indicated by the Kaggle uploader of the dataset, indicating possible instability in this regard.

3. The dataset was posted to Kaggle under a CC By 4.0 license, meaning that it can be freely shared and adapted as long as proper attribution is given and no additional legal restrictions are added. I believe I have followed these terms by adapting it for my Github repository with attribution, and would keep the same license and terms.

4. Human subject considerations: Though all of the data included pertains to humans, there should not be any issues regarding human subject safety or privacy. The data is extremely broad – at the national level – and records general population demographics rather than any personal or potentially invasive details. The data in itself should be anonymous, and any additional steps to obscure possible personal details seem to have been undertaken already by the original researcher, the World Bank.