WorldBank_WDI_Visulization

A Data Management Plan created using DMPonline

Creator: Craig Lincoln

Affiliation: Other

ORCID iD: https://orcid.org/0000-0003-3239-0114

Funder: European Commission (Horizon 2020)

Template: Horizon 2020 DMP

Project abstract:
Create an interactive visualization to explore World Bank - development indicators dataset.

Last modified: 20-04-2020

WorldBank_WDI_Visulization - Initial DMP

1. Data summary

Provide a summary of the data addressing the following issues:

- **State the purpose of the data collection/generation**
- **Explain the relation to the objectives of the project**
- **Specify the types and formats of data generated/collected**
- **Specify if existing data is being re-used (if any)**
- **Specify the origin of the data**
- **State the expected size of the data (if known)**
- **Outline the data utility: to whom will it be useful**

INPUT DATA

The dataset is the World Bank's - world development indicators which "*The World Development Indicators is a compilation of relevant, high-quality, and internationally comparable statistics about global development and the fight against poverty. The database contains 1,600 time series indicators for 217 economies and more than 40 country groups, with data for many indicators going back more than 50 years.*" The purpose of the project is to enable the user to find indicators that correlate with an indicator of interest (in focus) using an interactive visual tool. This project uses a downloadable csv version of the data but could be extended to streaming via the available API. The downloaded csv is around 60 Mb but this will increase with time as data is constantly added. The download includes 6 separate files:

- WDICountry-Series.csv - 7979 rows, 3 columns -"CountryCode","SeriesCode","DESCRIPTION",
- WDICountry.csv 270 rows, 30 columns - "Country Code","Short Name","Table Name","Long Name","2-alpha code","Currency Unit","Special Notes","Region","Income Group","WB-2 code","National accounts base year","National accounts reference year","SNA price valuation","Lending category","Other groups","System of National Accounts","Alternative conversion factor","PPP survey year","Balance of Payments Manual in use","External debt Reporting status","System of trade","Government Accounting concept","IMF data dissemination standard","Latest population census","Latest household survey","Source of most recent Income and expenditure data","Vital registration complete","Latest agricultural census","Latest industrial data","Latest trade data",
- WDIData.csv - 377257 rows, 64 columns - "Country Name","Country Code","Indicator Name","Indicator Code","1960","1961","1962","1963","1964","1965","1966","1967","1968","1969","1970","1971","1972","1973","1974","1975","1976","1977","1978","1979","1980","1981","1982","1983","1984","1985","1986","1987","1988","1989","1990","1991","1992","
- WDIFootNote.csv - 561476 rows, 3 columns -"CountryCode","SeriesCode","Year","DESCRIPTION",
- WDISeries-Time.csv - 463 rows, 3 columns - "SeriesCode","Year","DESCRIPTION",
- WDISeries.csv - 1429 rows, 20 columns - "Series Code","Topic","Indicator Name","Short definition","Long definition","Unit of measure","Periodicity","Base Period","Other notes","Aggregation method","Limitations and exceptions","Notes from original source","General comments","Source","Statistical concept and methodology","Development relevance","Related source links","Other web links","Related indicators","License Type",

This is essentially self contained in that there are long descriptions of each indicator and how the data is sourced. The data itself can be found in WDIData.csv and are all numeric variables. For more details one can refer to the World Bank Glossary (https://databank.worldbank.org/metadataglossary/all/series)

Intermediate data files:

- corr_matrix_sign.csv - <indicator name>, <indicator name> - sign of Pearsons correlation coefficient represented as +/-1 (integer)
- corr_matrix.csv - <indicator name>, <indicator name> - sign of Pearsons correlation coefficient (float)
- wdi_code_breakdown.csv - 1429 rows, 23 columns -'Series Code', 'Topic', 'Indicator Name', 'Short definition', 'Long definition', 'Unit of measure', 'Periodicity', 'Base Period', 'Other notes', 'Aggregation method', 'Limitations and exceptions', 'Notes from original source', 'General comments', 'Source', 'Statistical concept and methodology', 'Development relevance', 'Related source links', 'Other web links', 'Related indicators', 'License Type', 'Unnamed: 20', 'indicator_prefix', 'topic_prefix'
- wdi_country_code_name.csv - 265 rows, 2 columns - 'Country Name', 'Country Code'
- wdi_indicator_code_name_topic.csv - 1430 rows, 3 columns -'Indicator Name', 'Indicator Code', 'topic_prefix'
- wdi_pivot.csv - <country name>, <indicator name> - mean of years 1990-present
- wdi_small.csv - 247104 rows, 32 columns - 'Country Name', 'Indicator Name', '1990', '1991', '1992', '1993', '1994', '1995', '1996', '1997', '1998', '1999', '2000', '2001', '2002', '2003', '2004', '2005', '2006', '2007', '2008', '2009', '2010', '2011', '2012', '2013', '2014', '2015', '2016', '2017', '2018', '2019'

The utility of the WDI dataset is obvious. This visualization attempts to further enrich the utility by allowing the relationships between different indicators to be explored. Thus, making it useful to users who wish to explore causal versus symptomatic effects on an indicator of interest.

ENVIRONMENT:

- Data was read, processed, manipulated and displayed using python and Plotly Dash.
- The GitHub repo (https://github.com/clincolnoz/WorldBank_WDI_Visulization) contains a README.md with instructions on how to use the data and setup the environment using requirements.txt.

2. FAIR data

2.1 Making data findable, including provisions for metadata:

- **Outline the discoverability of data (metadata provision)**
- **Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?**
- **Outline naming conventions used**
- **Outline the approach towards search keyword**
- **Outline the approach for clear versioning**
- **Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how**

The project DOI is: https://zenodo.org/badge/latestdoi/234558640

The WDI DOI is: http://doi.org/10.5257/wb/wdi/2019Q4

The the project includes:

- documentation/metadata.xml (project description)
- documentation/metadata_WDI.xml (rdf of WDI dataset)
- documentation/WorldBank_WDI_Visulization-DMP.pdf (DMP)
- documentation/WorldBank_WDI_Visulization-maDMP.json (DMP)
- project file structure in README.md
- project is linked to author maintainer ORCID

As mentioned in the summary the dataset is largely self contained with extensive descriptions of the sources of the data and what each data series measures (See WDICountry-Series.csv, WDICountry.csv, WDIFootNote.csv, WDISeries-Time.csv and WDISereis.csv). Additionally, the data is freely available and well documented and the World Bank also maintains a glossary. The data specifically used by the project is available in the GitHub repository.

2.2 Making data openly accessible:

- **Specify which data will be made openly available? If some data is kept closed provide rationale for doing so**
- **Specify how the data will be made available**
- **Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?**
- **Specify where the data and associated metadata, documentation and code are deposited**
- **Specify how access will be provided in case there are any restrictions**

- The data source is given in the README.md and code is available at https://github.com/clincolnoz/WorldBank_WDI_Visulization (and https://zenodo.org/badge/latestdoi/234558640)
- Usage instructions are included in README.md
- Package requirements are included in requirements.txt
- WDI dataset licensed CC-BY 4.0
- Licensing (MIT) is included in LICENSE.txt and documentation/metadata.xml
- Project file structure included in README.md

2.3 Making data interoperable:

- **Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.**
- **Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?**

- The source data is csv tables (schema provided in summary)
- Preprocessing is done using Pandas and intermediates are stored as csv tables (schema outlined in summary)
- This DMP can be found in documentation/WorldBank_WDI_Visulization-DMP.pdf

2.4 Increase data re-use (through clarifying licenses):

- **Specify how the data will be licenced to permit the widest reuse possible**
- **Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed**
- **Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why**
- **Describe data quality assurance processes**
- **Specify the length of time for which the data will remain re-usable**

Standard MIT with limits

see https://github.com/clincolnoz/WorldBank_WDI_Visulization/blob/master/LICENSE.txt

3. Allocation of resources

Explain the allocation of resources, addressing the following issues:

- **Estimate the costs for making your data FAIR. Describe how you intend to cover these costs**
- **Clearly identify responsibilities for data management in your project**
- **Describe costs and potential value of long term preservation**

GitHub is free!

4. Data security

Address data recovery as well as secure storage and transfer of sensitive data

Relying on the Internet, World Bank, GitHub and Zenodo.org for backup and recovery
The project DOI is: https://zenodo.org/badge/latestdoi/234558640

5. Ethical aspects

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

There is no attempt to identify nor mitigate any ethical aspects present in the World Bank - World Development Indicator dataset.
See https://www.worldbank.org/en/about/unit/ethics_and_business_conduct for the World Bank's governance.

6. Other

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

N/A