**Data Processing Across Domains Using Shared Libraries and Best Practices**

**Best Practice Guidelines and Recommendations**

**Introduction**

Many of the challenges being by the world can only be analyzed if researchers can access and use data from different domains. Such data are usually available from different sources and come with different characteristics. The opportunities and challenges associated with assembling data from a number of domains were examined through the use of a case study.

**Objectives**

The objectives of this exercise are to determine what types of challenges researchers could encounter if they were to analyze a global issue by accessing data with varying characteristics from different sources. By working their way through a case-study using typical data, the group was about to identify data issues, ways of solving them and identifying best practices as well as opportunities for improvement.

 **Cross Domain Case Study to Illustrate Data and Metadata Challenges and Opportunities**

Globally, the infant mortality rate has decreased from an estimated rate of 65 deaths per 1000 live births in 1990 to 29 deaths per 1000 live births in 2017. Annual infant deaths have declined from 8.8 million in 1990 to 4.1 million in 2017. However, the risk of a child dying before their 1st birthday was highest in the WHO African Region (51 per 1000 live births) was over six times higher than that in the WHO European Region (8 per 1000 live births). Policy makers would like to determine the underlying causes of such a disparity and to identify which factors could be influenced to reduce this number of countries with the highest infant mortality.

In our scenario, researchers decided to analyze mortality rates for infants under 5 years of age. This age range was chosen as the data for this range were the most readily for a wide range of countries. **[tailoring analysis to available data is a choice that researchers often have to make]**

***Placeholder: Note on methodology. Scenario where we will demo 3 versions. Version 1 …the approach Erol and group 5 saw/did. Version 2, better lookup etc, version 3 is with automation*.**

**The Data Sources**

Data relevant to this issue can be gleaned from a number of sources from different suppliers and formats in micro data files as well as tables of aggregate numbers.

* Data on infant mortality by country was obtained UNICEF. These data are expressed as rates per 1000 population. <https://data.unicef.org/topic/child-survival/under-five-mortality/>
* Data about family characteristics come from a survey level (micro) dataset from the Demographic and Health Survey (DHS). These data described the family situation for the employment and HIV status of the parents. This survey file was downloaded from the data supplier **[link]** and were prepared for analysis in the Python software. A detailed description of this activity follows below.
* Data on the Gross Domestic Product (GDP) of the countries are aggregate data and were obtained from the World Bank. <https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD>
* Land use.....
* *Air quality was identified as a possible contributing factor but since it was only available for specific locations (specific cities) it was not included in the global analysis. Space for it was left available in the table below in case there was an opportunity to extend this analysis below the national level.*
* The analytic data components are shown in the table below.

**Infant Mortality in the context of various Indicators by country**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Country** | **Infant Mortality** | **Employment status of Parents** | **HIV Prevalence** | **National GDP** | **Land Use** |
| **Angola** |  |  |  |  |  |
| **Burundi** |  |  |  |  |  |
| **Ethiopia** |  |  |  |  |  |
| **….** |  |  |  |  |  |

**Processing and Cleaning the DHS for a Single Reuse *(to be reviewed)***

Within a project, once external data have been identified and acquired there may need to be extensive data processing in order to make the data suitable for the project’s purpose. In some cases, this might mean cleaning operations like making sure the values for variables fall into the defined value domains. In other cases, a variable might not be in the correct form for input into an analytic procedure. Similarly, the structure of a dataset might need to be rearranged, such as a transposition from a tall structure to a wide structure.

A research group might benefit by developing a library of transformation operations, either as a set of descriptions or as a set of machine actionable scripts. For our example study we looked through a script of about 2000 lines of Python code used to clean DHS data for a machine learning study. We attempted to classify these operations as follows. Some of these operations are generic, being applicable to many uses. Others are specific to the needs of the machine learning application of our use case.

### **Sample operations from the DHS use case**

#### **Generic**

* Variables with a constant value are dropped
* Variables that completely duplicate another variable are dropped
* Cases with specific (substantive or sentinel) values are dropped
* Codes with misspellings are mapped to a single value
* Case of codes is regularized
* Variables are renamed with semantic names (Identifying and representing top and bottom coding (Number of injections in the last 12 months = »90+ »)
* Identifying and dealing with outliers

#### **Specific purpose driven**

* Some nominal variables are dummy coded
* Variables with more than some percentage missing are dropped
* Top and bottom coding is transformed using a fixed value or reference value (e.g. person’s age)
* Nuanced sentinel values are mapped to a common value (e.g. Stata missing values to R, Python NaN or NA)
* Some ordinal variables are transformed into interval variables
* Collapse categories for some variables (e.g. « catholic, roman Catholic)  « Catholic »
	+ This is best done using some international classification scheme
* Value mappings may follow complex rules, for example if fewer than five variables in a set are originally missing (NA) then recode to .1, but if all variables in the set are NA then do not recode)
* Dropping variables (survey paradata may not be relevant)
* Drop variables with low variance….(this is a choice depending on analysis requirements)
* Imputing missing values

We noted that a library of these operations could be built over time by a research group and used to train staff or even to automate at least some of the data processing.

### **Mappings between value domains**

One of the issues that this exercise revealed was that the set of country codes used by DHS did not match the ISO 3166-1 country codes. In particular as seen in the table below Burundi and Namibia have different codes in the two sets. Good practice would be for data shared from our example to recode the DHS codes to ISO codes. It should be noted that these codes would change over time as countries change. The set of codes to be shared should be time stamped.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **DHS** |  | **ISO 3166-1** |
| Angola | AO | --> | AO |
| Burundi | BU | --> | BI |
| Ethiopia | ET | --> | ET |
| Lesotho | LS | --> | LS |
| Malawi | MW | --> | MW |
| Mozambique | MZ | --> | MZ |
| Namibia | NM | --> | NA |
| Rwanda | RW | --> | RW |
| Zambia | ZM | --> | ZM |
| Zimbabwe | ZW | --> | ZW |

# **Populating a Table**

As an exercise we decided to populate a table from the DHS data with these concepts for rows and columns.

*To be reviewed in the context of the larger study.*

**HIV Cases**

|  |  |  |
| --- | --- | --- |
| Country | Male | Female |
| Angola |   |   |
| Burundi |   |   |
| Ethiopia |   |   |
| Lesotho |   |   |
| Malawi |   |   |
| Mozambique |   |   |
| Namibia |   |   |
| Rwanda |   |   |
| Zambia |   |   |
| Zimbabwe |   |   |

### **Codes for Sex**

One of the team knew that the Australian Bureau of Statistics had a classification scheme for sex. A google search for “abs sex classification” yielded the page:

[ttps://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/1200.0.55.012Main%20Features212016](https://www.abs.gov.au/ausstats/abs%40.nsf/Latestproducts/1200.0.55.012Main%20Features212016)

Table 1 below describes the category codes, labels, and definitions of the sex classification and the gender classification.

**TABLE 1. THE SEX AND GENDER STANDARD CLASSIFICATIONS AND CODE STRUCTURES**

|  |
| --- |
|  |
| Preferred Code | Alternate Code | Label | Definition |
|  |
| **The Sex Standard Classification** |
|  |
| 1 | M | Male | Persons who have male or predominantly masculine biological characteristics, or male sex assigned at birth. |
| 2 | F | Female | Persons who have female or predominantly feminine biological characteristics, or female sex assigned at birth. |
| 3 | X | Other | Persons who have mixed or non-binary biological characteristics (if known), or a non-binary sex assigned at birth. |
|  |
| **The Gender Standard Classification** |
|  |
| 1 | M | Male  | Adults who identify themselves as men, and children who identify themselves as boys. |
| 2 | F | Female | Adults who identify themselves as women, and children who identify themselves as girls. |
| 3 | X | Other | Adults and children who identify as non-binary, gender diverse, or with descriptors other than man/boy or woman/girl. |
|  |

These were hand coded into our application.

### **Codes for Countries**

The DHS data has the country categories coded, but we wanted to ensure standard codes for the countries and sex categories. The original Stata file represents the country variable as a three character string like “BU7”. The metadata for the DHS files just describes this as a three character string, without any identification of the country associated with the first two characters. The numeric character is the vintage of the dataset.

We wanted to find a source for standard country codes, preferably ISO codes. We first searched Google for “iso country code”. We found this unofficial source:

<https://en.wikipedia.org/wiki/List_of_ISO_3166_country_codes>

and could have web scraped it but we wanted to find an open official source, either as a table or API. ISO codes are only available for a payment.

After considerable exploration we found there is a crosswalk for country codes available at:

<http://api.dhsprogram.com/rest/dhs/countries>

The Entry for Burundi is:

Africa","WHO\_CountryCode":"BI","FIPS\_CountryCode":"BY","ISO2\_CountryCode":"BI","ISO3\_CountryCode":"BDI","RegionOrder":13,"DHS\_CountryCode":"BU","CountryName":"Burundi","UNICEF\_CountryCode":"BRD","UNSTAT\_CountryCode":"BDI","RegionName":"Sub-Saharan Africa"},{"UNAIDS\_CountryCode":"KHM","SubregionName":"Southeast

Note that the ISO code for Burundi is “BI” but DHS uses “BU”. There is no country associated with an ISO code of “BU”. This site provides the information necessary to develop the mapping from DHS codes to ISO codes.

Gathering the ISO codes and the codes used in DHS reveals that the following transformation would be necessary to map the DHS codes into ISO codes. This is a duplicate

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **DHS** |  | **ISO 3166-1** |
| Angola | AO | --> | AO |
| Burundi | BU | --> | BI |
| Ethiopia | ET | --> | ET |
| Lesotho | LS | --> | LS |
| Malawi | MW | --> | MW |
| Mozambique | MZ | --> | MZ |
| Namibia | NM | --> | NA |
| Rwanda | RW | --> | RW |
| Zambia | ZM | --> | ZM |
| Zimbabwe | ZW | --> | ZW |

Once the codes and a mapping were developed, a team member wrote R code to collect the data to populate the desired table.

*Insert Steve’s API*

- <http://rpubs.com/stevenmce/DagstuhlGroup5_R_Example1_NOW>

**Challenges and Opportunities**

* The state of the data file selected for analysis.
	+ Feedback and guidelines for data suppliers on desirable characteristics of dissemination files.
	+ Create libraries of transformation operations, either as a set of descriptions or as a set of machine actionable scripts for use in file cleaning
	+ Share libraries of transformation operations or clean setup files among a group of users. E.g., with a Python user group, list serves such as IASSIST??
* Classifications from other organizations?
	+ Xccx
	+ Xcv

Recommendations

* + best practices *(review and insert Dan’s API)*
	+ Opportunities for shared practices
		- A Google search for “DataCleaning Checklists” reveals a number of lists of data cleaning operations. A repository of these, even at just the descriptive level, would encourage the creation of cleaner shared data. Among these practices would be easily findable references to standard codelists. Even better would be to encourage use of elements by reference to the common repository. Harmonizing variables from different sources is much easier if they both refer to the same value domain rather than duplicating it.
		- ….
	+ Opportunities for automation
	+ Key areas for action