Unabridged Taxonomies of Data Wrangling in Computational Journalism

Here we present complete versions of our two taxonomies of data wrangling in computational journalism: *Actions* and *Processes*. We use shortcodes to refer to longer descriptions of open and axial codes in the paper, follow this naming convention:

<1>.<2>.<3>.<4>.<5>

<1>: *A* for Action and *P* for Process

<2>: The first character of the code, capitalized

<3>: Letters a-z, lowercase

<4>: Arabic numerals, 1 - 9

<5>: Roman numerals, lowercase, i - x

Actions Taxonomy

The *Actions* taxonomy details individual data wrangling steps made by journalists.

A.I. **Import**: How raw data is introduced into the wrangling environment

- A.I.a. **Fetch**: Data is retrieved from a source external to the wrangling environment
  - A.I.a.1. **Extract Data from PDF**: Using a data extraction tool, such as Tabula, to parse tables inside PDF documents
- A.I.b. **Create**: Data is created inside the wrangling environment
  - A.I.b.1. **Construct Data Manually**: The data is either copy-and-pasted or values are created manually
  - A.I.b.2. **Generate Data Computationally**: Using data with values generated programmatically
  - A.I.b.4. **Impute Missing Data**: Replace missing data values either manually, through data entry, or systematically through an R functions such as lag.
- A.I.c. **Load**: Data resides on the local disk and is loaded into the environment

A.C. **Clean**: The process of removing incorrect, incomplete, inaccurate, misformatted or otherwise corrupt observations, variables, and values within a dataset.

- A.C.a. **Remove**: Approaches to cleaning data that involve removing observations, variables, and values

A.C.a.1. **Deduplicate**: Remove duplicate observations

A.C.a.2. **Remove Non-Data Rows**: Remove notes and comments that are not observations

A.C.a.3. **Remove Incomplete Data**: Drop observation if it contains incomplete values, often denoted as NA or Null

A.C.b. **Replace**: Approaches to data cleaning that involve replacing observations, variables, and values

A.C.b.1. **Replace NA Values**: Journalists replace NA values of a variable with other values. NA values can be denoted in various ways

A.C.b.2. **Edit Values**: Existing data values are incorrect, but not NA, and must be changed by the journalist

A.C.b.3. **Resolve Entities**: Resolving the issue of different categorical values for the same entity. This wrangling action is particularly suited towards fixing common data issues such as misspellings, inconsistent date formats, and name ordering.

A.C.b.4. **Standardize Categorical Variables**: Make levels conform to some set of rules, such as replacing whitespace for underscore, trimming whitespace, etc

A.C.b.5. **Scale Values**: Operations that apply some mathematical operation to a quantitative variable in the spirit of fixing data errors. For example, quantitative data may be in the millions and only display significant digits.

A.C.c. **Reformat**: Wrangling operations that modify the table entry's appearance or style, but not value

A.C.c.1. **Format Values**: Operations that change value appearance, but not the underly variable type: changing case, specifying date format, rounding floats

A.C.c.2. **Canonicalize Variable Names**: Operations that change column names

A.M. **Merge**: Operations that combining multiple datasets

A.M.a. **Union Datasets**: Combining multiple datasets with identical variables into one dataset

A.M.b. **Inner Join**: Take the intersection of two datasets on a shared key variable

A.M.c. **Supplement**: The variables of one dataset are supplemented with the variables of another dataset

A.M.c.1. **Outer Join**: Retain observations with no corresponding match in the dataset being joined upon

A.M.c.2. **Full Join**: Retain observations with no corresponding match in either dataset

A.M.c.3. **Concat Parallel Datasets**: Join two datasets by position, without specifying a joining key

A.M.d. **Cartesian Product**: Create a new dataset by the unique pairing of each key in their respective datasets

A.M.e. **Self Join Dataset**: Create a new dataset by joining it with itself

A.P. **Profile**: Operations the inspect the state of the data during wrangling

A.P.a. **Run a Test**: Audit the data by constructing a pass or fail scenario

A.P.a.1. **Report Rows With Column Number Discrepancies**: Check the number of columns or rows between tables

A.P.a.2. **Test for Equality**: Test if two data structures are exactly the same

A.P.a.3. **Test for Null Values**: Test the results of a calculation against different methods/packages

A.P.a.4. **Validate Data Quality with Domain-Specific Rules**: Test with a domain-specific rule for the data, such as checking if the average temperature is higher than the maximum recorded value

A.P.b. **Check Results**: Output the dataset for review

A.P.b.1. **Peek at Data**: Display the first *n* observations, or take a random sample, with all variables of the dataset

A.P.b.2. **Inspect Data Schema**: Check the data types of columns

A.P.b.3. **Select Rows with Missing Values**: Inspect the dataset for observations with a missing value, often denoted as NA

A.P.b.4. **Check for NAs**: See if any observations have NA values

A.P.b.5. **Visualize Data**: Employ any kind of data visualization, including tables

A.P.c. **Summarize Dataset**: Summarize the dataset numerically

A.P.c.2. **Count Unique Values**: Report the number of unique values in one or more variables

A.P.c.3. **Describe Statistically**: Generate descriptive statistics of the dataset, such as central tendency, dispersion, or distribution shape

A.D. **Derive**: Expand upon the original dataset without integrating another dataset

A.D.a. **Detrend**: Remove the secular effect from a variable; these are not considered data cleaning operations because values are not erroneous

A.D.a.1. **Adjust for Inflation**: Remove the effect of price inflation from data

A.D.a.2. **Compute Index Number**: Calculate the change in a variable over time

A.D.a.3. **Adjust for Season**: Adjust a variable to compensate for seasonal effect

A.D.b. **Consolidate Variable Values**: Map a set of unique values to a smaller set, which is different from entity resolution

A.D.b.1. **Bin Values**: Consolidate a quantitative variable into a smaller set of ordinal data

A.D.b.2. **Combine Categorical Values**: Consolidate the levels of a categorical variable into a smaller set of levels.

A.D.c. **Generate Unique Identifiers**: Attempt to create unique identifiers

A.D.c.1. **Generate Observation Identification**: Produce unique identification for each observation

A.D.c.1.i. **Create Soft Key**: Keys not guaranteed to be unique per observation

A.D.c.1.ii. **Create a Unique Key**: Keys are guaranteed to be unique per observation

A.D.c.2. **Generate Dataset Identification**: Add a table identification value as a variable for all observations

A.D.d. **Subset the Dataset**: Reduce the size or complexity of the actively wrangled dataset

A.D.d.1. **Remove Variables**: Specify which variables to remove or retain from a dataset

A.D.d.2. **Remove Observations**: Specify which observations to remove or retain from a dataset

A.D.d.2.i. **Trim by Date Range**: Remove based on observations inside or outside a range of dates

A.D.d.2.ii. **Trim by Geographic Area**: Remove based on observations inside or outside the geographic region

A.D.d.2.iii. **Trim by Quantitative Threshold**: Remove based on observations above, below, equal to, or not equal to a quantitative value

A.D.d.2.iv. **Trim by Categorical Value**: Remove based on observations that do or do not contain specific a specific value

A.D.e. **Formulate a Performance Metric**: Calculate a quantitative variable

A.D.e.1. **Assign Ranks**: Order observations explicitly as a variable

A.D.e.2. **Standardize Variable**: Measure deviation from "normal," such as z-scores

A.D.e.3. **Figure a Rate**: Calculate a normalized rate to provide a baseline for comparison

A.D.e.4. **Calculate Change Over Time**: Calculate percentage change over time

A.D.e.5. **Calculate Spread**: Calculate the difference between two values or rates

A.D.e.6. **Domain-Specific Performance Metric**: Calculate a domain-specific metric

A.D.e.7. **Get Extreme Values**: Calculate the highest or lowest values in a variable

A.T. **Transform**: Create or revise table variables based on existing variables, without *integrating* other tables

A.T.a. **Reshape**: Change the table's structure, without summarizing any data.

A.T.a.1. **Transpose**: Change places between rows and columns within a table or Matrix

A.T.a.2. **Cross Tabulate**: Create a pivot table or crosstab

A.T.a.3. **Spread Table**: Expand two columns of key value pairs into multiple columns

A.T.a.4. **Gather**: Collapse table into key value pairs

A.T.a.5. **Create a Flag**: Spread a categorical variable into multiple boolean variables.

A.T.b. **Modify Variables**: Change properties of variables within a dataset

A.T.b.1. **Parse Variable**: Separate variables into multiple new variables using position or regular expressions

A.T.b.2. **Consolidate Variables**: Combine two different variables into one composite variable

A.T.b.3. **Replace Variable Levels**: Change the value of a level in a categorical variable to another value

A.T.c. **Summarize**: Aggregate observations to summarize a phenomenon; we consider this a structural change as it effectively coarsens the dataset

A.T.c.1. **Group By Variable**: Group or partition by the levels of one or more categorical variables

A.T.c.2. **Aggregate**: Aggregate quantitative values using functions such as sum, mean, median, or count

A.T.c.3. **Rolling Window Calculation**: Perform rolling-window aggregation

A.T.d. **Sort**: Order observations implicitly by position within a data structure

A.E. **Export**: Export the results of data wrangling either by writing results to disk or return data from a function

Process Taxonomy

The process taxonomy consists of the paper authors' interpretations of the processes that occur during data wrangling.

P.S. **Source**: Codes that describe how the raw data was obtained by journalists

P.S.a. **Collect Data**: Journalists are the initial data collector

P.S.a.1. **Collect Raw Data**: The journalist collected the raw data themselves.

P.S.a.2. **Freedom of Information Data**: Data that was obtained via FOI/FOIA requests

P.S.b. **Acquire Data**: Journalists acquired data from another party

P.S.b.1. **Use Previously Cleaned Data**: Data that originated from a colleague

P.S.b.2. **Use Public Data**: Includes open-source datasets, datasets on Wikipedia, etc

P.S.b.3. **Use Academic Data**: Use data collected from an academic study

P.S.b.4. **Use Non-Public, Provided Data**: Use data that is not publically available

P.S.b.5. **Use Open-Government Data Portal:** Data is publicly available on civic data portals.

P.S.b.6. **Use Another News Orgs Data:** Use another news organization's data

P.S.b.7. **Use a Colleague's Data:** A dataset was provided by another journalists

P.W. **Workflow**: Codes pertaining to how the wrangling workflow is built

P.W.a. **Annotations**: Adding comments or notes in Markdown that explain what the journalists are doing

P.W.b. **Computational Process:** Codes that demonstrate computational thinking on the part of the journalist

P.W.b.1. **Construct a Subroutine**: A set of instructions grouped together to be performed multiple times

P.W.b.2. **Construct Data Pipeline**: An instance where one script is designed to handle multiple data sources. Often journalists construct subroutines and loops.

P.W.c. **Toggle Operation**: Ensuring that some code segments are not always run, such as by commenting out lines of code

P.C. **Cause**: Based on the final output and comments, why does it seem like this data needs to be wrangled?

P.C.a. **Downstream Input**: Output from wrangling will be input into some other program

P.C.a.1. **Wrangle Data for Graphics**: Data needs to be formatted in order to be visualized in an article, including datasets.

P.C.a.2. **Wrangle Data for Model**: Data is being wrangled in order to create a model, whether the main point of the piece is for prediction or classification

P.C.a.3. **Create New Datasets**: These raw datasets are being wrangled in order to create a new dataset

P.C.a.3.i. **Combine Periodic Data**: Combine many separate datasets published over time into one dataset

P.C.a.3.ii. **Merge Seemingly Disparate Datasets**: When a notebook largely constitutes combining seemingly unrelated datasets

P.C.a.3.iii. **Geolocate Dataset Records**: Pairing data with GIS info

P.C.a.4. **Generate High-Level Summary**: Data of individual observations is aggregated in an attempt to find some meaningful structure or patterns

P.T. **Themes**: General themes for how data objects are transformed throughout the wrangling process

P.T.a. **Divide and Conquer**: Instances where the data wrangling processes separates one objects into smaller components

P.T.a.1. **Split, Compute, and Merge**: First, the journalist partitions a single data frame into multiple, separate data frames. Then, often identical computations are run on all the data frames. Finally, the multiple data frames are consolidated into one data frame again

P.T.a.2. **Split and Compute**: One dataset is split into two or more and identical computations are applied to each dataset

P.T.b. **Join Aggregate**: When aggregated statistics about a dataset are added to the datasets as a variable, either column-wise or row-wise (as with adorn_totals in R)

P.T.c. **Create a Frequency Table**: A table the displays the frequency of categorical variables within a column

P.T.d. **Trim Fat**: Trim the fat refers to when large amounts of observations or variables are removed from the dataset early in the wrangling processes, if not as the first step of wrangling. We infer that these sections are irrelevant to further analysis.

P.T.e. **Align Variables**: Modifying dataset variables to match each other, often prior to merging datasets.

P.A. **Analysis**: Kinds of analysis data journalists need to wrangle data to perform

P.A.a. **Interpret Model:** Analyze features from a model such as linear regression or classification trees.

P.A.b. **Compare Groups**: The end analysis is just comparing different groups by a common metric

P.A.c. **Identify Extreme Values**: Identify values that are at the ends of the range, but not strictly outliers

P.A.d. **Show Trend Over Time**: Analysis consists of showing how values change over time

P.A.e. **Calculate a Statistic**: Calculate a single value from a dataset, such as number of records

P.A.f. **Count the Data**: Analysis involves count-based metrics on the datasets including percentages, with optional filtering and aggregation

P.A.g. **Lookup Table Values**: Analysis consists looking up values in a table

P.A.h. **Examine Relationship**: Analysis consists of examining the relationship between different phenomena

P.A.i. **Explain Variance**: This can be done via PCA

P.A.j. **Search for Clusters**: Look for groups within the data where its presence, or lack thereof, is significant

P.A.k. **Perform Network Analysis**: Journalists perform any kind of network analysis, such as finding all nearest neighbors in the network

P.A.l. **Explore Dynamic Network Flow**: (Network analysis) explore the flow between different nodes in the graph, such as migration between cities

P.A.m. **Create Lookup Table**: Make a table with two columns to map from one value to another

P.A.n. **Aggregate Join**: Aggregating a table and then joining those results to the original table

P.M. **Management**: General strategies journalists for managing data within the wrangling environment

- P.M.a. **Object Persistence**: How do journalists regard previous versions of datasets after applying transformation functions?
  - P.M.a.1. **Data Evolves**: Data and objects are overwritten and replaced during the wrangling process
    - P.M.a.1.i. **Variable Replacement**: The output of any column calculation is reassigned to an existing column
    - P.M.a.1.ii. **Temporary Joining Column**: When a key for joining two datasets is created and deleted immediately after the join
    - P.M.a.1.iii. **Refine Table**: Dataset refinement refers to when a table is subset in place, a new object is not created in the environment
- P.M.b. **Data Quality**: How journalists proceed when data may be incomplete, erroneous, or otherwise not 100% clean
  - P.M.b.1. **Set Data Confidence Threshold**: Removes rows where a quantitative value is less than, greater than, or not equal to a numeric value
  - P.M.b.2. **Tolerate Dirty Data**: Analysis continues despite clear data quality issues

P.P. **Pain Points**: Areas where journalist seem/could be frustrated in the wrangling process

- P.P.a. **Fix Incorrect Calculation**: Calculations in the data are incorrect and the journalist must recalculate them
- P.P.b. **Repetitive Code**: Instances where code is repetitively copied and pasted
- P.P.c. **Make an Incorrect Conclusion**: Instances where the journalist has made an incorrect conclusion about the data
- P.P.d. **Post-Merge Clean Up**: Pain points that come from the result of merging two datasets together
  - P.P.d.1. **Resort after Merge**: When a sort has to be re-done because a merge ruining the pre-merged order
  - P.P.d.2. **Fill in NA Values After an Outer Join**: As outer joins do not drop non-matching rows, those values have NA
  - P.P.d.3. **Lossy Join**: When data is lost after integrating two tables
  - P.P.d.4. **Remove Duplicate Variables**: Two tables may have duplicate variables and duplicate variables need to be removed
- P.P.e. **Post-Aggregation Clean Up**: Pain points that come from the result of grouping a table
  - P.P.e.1. **Data Loss from Aggregation**: When table columns are lost because they were dropped form resulting dataset due to not being relevant in aggregation
  - P.P.e.2. **Silently Dropping Values After Groupby**: Values other than those being grouped and calculated upon are lost in a group-by operation
- P.P.f. **Data too Large for Repo**: Raw data cannot be included because files are too large

P.P.g. **Schema Drift**: When the schema of a perennially published datasets varies from edition to edition

P.P.h. **Data Type Shyness**: Users often seem to avoid using built in data types