Table Scraps: An Actionable Framework for Multi-Table Data Wrangling from an Artifact Study of Computational Journalism

Stephen Kasica, Charles Berret, Tamara Munzner

INTRODUCTION

There exists abundance of literature on data wrangling in the context of enterprise data analysis. However, little is known about the specific operations, processes, and pain points of data wranglin in journalism. To better understand the needs of this unique group, we conduct a technical observation study of journalists' code repositories, *in the wild*. The research questions addressed in this paper are:

Q1 What are the wrangling practices of journalists?

Q2 Which practices align with or diverge from existing characterizations?

Q3 How to re-characterize wrangling to match the observed practices?

In answering these questions, we provide the following primary contributions:

C1 Two detailed and cross-cutting taxonomies of data wrangling in computational journalism, for actions and for processes

C2 A concise, actionable framework for general multi-table data wrangling

CRITICAL INCIDENTS

We presents two usage scenarios taken from the pool of 50 repos in our technical observation study to highlight specific critical incidents:

MULTI-TABLE WRANGLING FOR THE WIN: One story of success when wrangling multiple tables

- *The Los Angeles Times* investigates districting water usage following a multi-year, state-wide drought
- Successfully reshapes raw data unfixable by single-table methods

MULTI-TABLE WRANGLING GONE AWRY: A cautionary tale of pitfalls and pain points when wrangling multiple tables

- *BuzzFeed News* explores 10 years of refugee resettlement data in the U.S.
- Journalists issues a editorial correction due to a data wrangling error

PROCESS

PHASE 1 (technical observation): Access digital artifacts such as source code produced by journalists to answer Q1

PHASE 1.1 (repo selection): Select relevant code repositories, resulting in a pool of 225 repos of data analysis by journalists

PHASE 1.2 (qualitative coding): Iterative process of open and axial coding data journalists analysis code, producing 50 coded repositories

PHASE 2 (literature search): Review relevant literature on wrangling multiple tables to assess Q2

PHASE 3 (reflective synthesis): Synthesize our taxonomy with related work to create a general multi-table framework for data wrangling

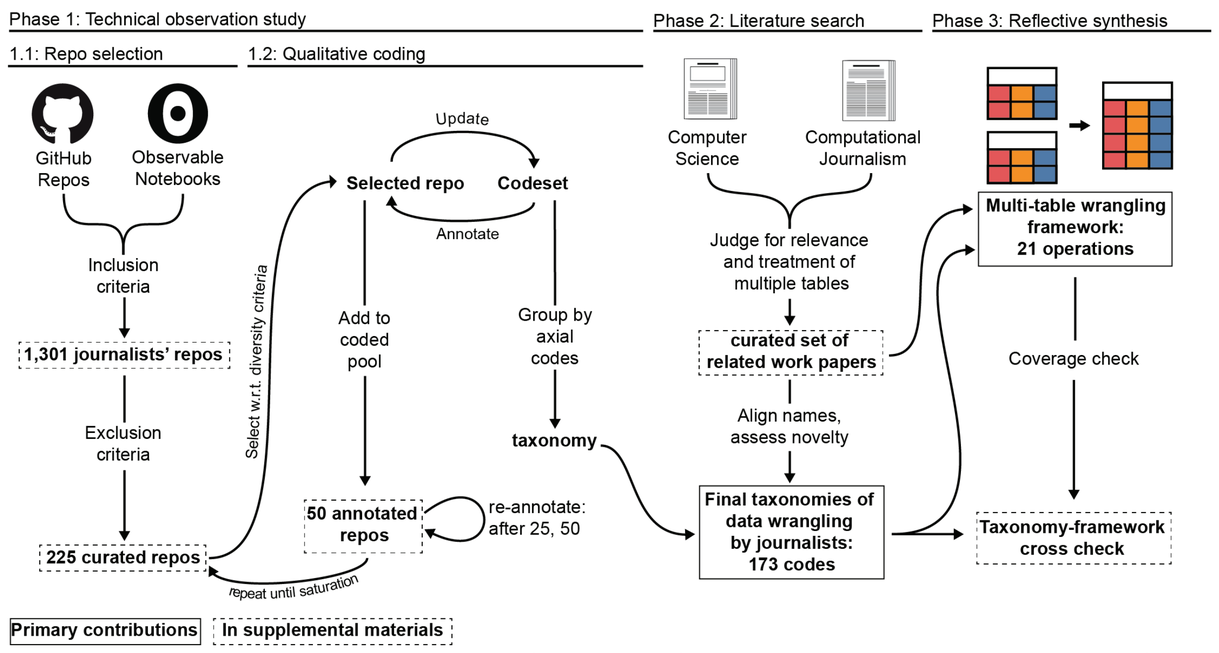


Fig. 1. Three-phase process: observation study of technical artifacts conducted through qualitative coding of journalist repos, resulting in two initial bottom-up taxonomies of 173 open and axial codes; literature search to align naming and assess novelty; reflective synthesis to create a concise top-down multi-table wrangling framework with 21 operations.

RESEARCH ARTIFACTS

C1: TAXONOMIES

We characterize data wrangling in computational journalism with two cross-cutting taxonomies: Actions and Process.

Actions	Process
Import	**Source**
Fetch	Collect Data
Create	Acquire Data
Load	**Workflow**
Clean	Annotations
Remove	Comp. Processes
Replace	Toggle Operations
Reformat	**Cause**
Merge	Downstream Input
Union Datasets	**Themes**
Supplement	Divide and Conquer
Cartesian Product	Trim Fat
...	Align Variables
Profile	Create Freq. Table
Run a Test	...
Check Results	**Analysis**
Summarize Dataset	Compare Groups
Derive	Count the Data
Detrend	...
Subset the Dataset	**Management**
Consolidate Variable	Object Persistence
...	Data Quality
Transform	**Pain Points**
Reshape	Merge Clean Up
Modify Variables	Aggregate Clean Up
Summarize	Schema Drift
Sort	Repetitive Code
Export	Data Type Shyness
	...

Table 1. Abridged version of our two descriptive taxonomies of data wrangling in computational journalism. The Actions taxonomy describes actions journalist made upon their data over the course of wrangling, and the Process taxonomy documents our interpretations of the journalists' processes.

C2: MULTI-TABLE FRAMEWORK

A concise, actionable framework for general multi-table data wrangling consists of 15 categories comprised of five operations on three data types.

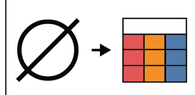
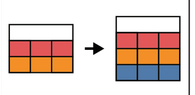
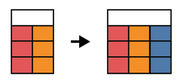
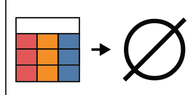
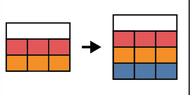
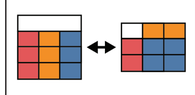
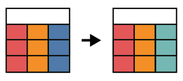
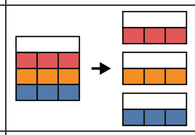
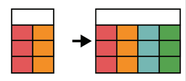
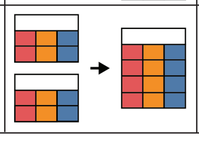
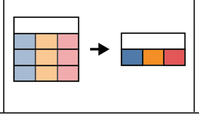
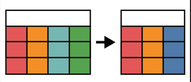
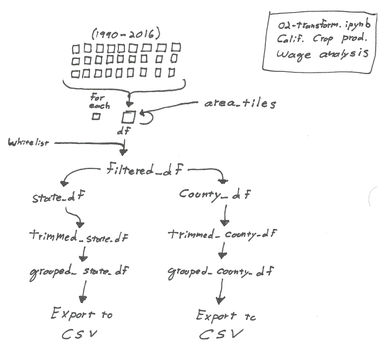
	Tables	Rows	Columns
Create 0 → 1			
Delete 1 → 0			
Transform 1 → 1			
Separate 1 → n			
Combine n → 1			

Table 2. Our Multi-table Framework reflects operations documented in our taxonomy that are without clear parallels in other work. This framework, the first to incorporate tables as first-class objects, will support future interactive wrangling tools for both computational journalism and general-purpose use.

FLOW DIAGRAMS

We sketch table-based flow diagrams visualizing the how raw data moves through the wrangling context when tables were used in complex ways.



- We sketch flow diagrams from 23 code repositories out of 50 (46%)
- Illustrative of our central finding: journalists often employ many tables in ways not addressed by related work

DISCUSSION

MANY TABLES

Journalists use and reuse many tables while wrangling data.

- One coded repo uses 38 tables objects over the course of combining and cleaning raw data
- Combining tables is common fare in modern newsrooms [Cohen et. al, 2011]

LAST-MILE WRANGLING

Journalists' raw data is relatively clean and well structured, but still requires much wrangling.

- Motivating examples of raw data in wrangling literature is barely structured
- Raw data comes in a structured format, but requires further tweaking to match the user's mental model

JOURNALISM MATTERS

Developing better tools for data wrangling would have the most impact on journalism that benefits society at large.

- Investigative journalism that holds public institutions rely heavily on data analysis [Hamilton, 2018]
- Data cleaning can take up to 80% of an analyst's time [Dasu and Johnson, 2003]

[kasica, tmm]@cs.ubc.ca
cberret@mail.ubc.ca
@swkasica, @cberret, @tamaramunzner
THE UNIVERSITY OF BRITISH COLUMBIA