

Codebook for:
Transparency, interactivity, diversity, and information provenance in everyday data journalism (Zamith, 2019)

1. Headline

What is the headline for this item?

- Open-ended: Copy and paste the headline

2. Number of Writers

How many individuals were credited with the production of this item?

- Open-ended: Enter the number of people credited with a byline.

3. Topic/Section

What is this story primarily about? You may use the section of the website or URL ('/politics/') as an indicator, but it doesn't have to be the determinant.

- 1 = Culture: News pertaining to art, literature, music, cinema, theatre, etc.
- 2 = Defense/National Security: Matters pertaining to defense and national security, including military operations and alliances, defense plans, the cost of military equipment or cost overruns, cyberattacks, etc.
- 3 = Economy: The distribution of goods and resources; trade; financial markets; changes in the availability of employment and/or workers' wages. Note that stories that focus on the economic implications of a policy or technology should be coded as 'economy.'
- 4 = Education: The efficacy of an educational program; the cost or value of attending school or university; the rankings of particular educational institutions.
- 5 = Health, Science, and Environment: The efficacy of certain medicines; the technological advances in a sector; the environmental impact of new technologies or industries.
- 6 = Judicial: Matters pertaining to legal rationales or background of lawyers and judges, including previous case histories for Supreme Court nominees, incidence of patent challenges in a district, and shifts in the political balance of courts.
- 7 = Policy: The setting of budgets at federal, state, and local levels; the development of tax plans; developing policies for combating climate change. Note that these stories should focus on the contents of policy in a more generic sense. If they focus on the economical implications, code as "economy"; social implications, as "society"; etc. Also note the distinction between "policy" and "politics," where the latter focuses on the strategic elements and the former on the output of that negotiation.
- 8 = Politics: Negotiations involving government bodies and/or officials; the production of laws; the 'horse race' among publicly elected officials and/or the political repercussions of a position; citizens' engagement with politics (e.g., protests); corruption and wrongdoing by officials.

- 9 = Sports: Salaries of sports players or coaches; comparing the performance of players or historicizing a performance; trade patterns between teams.
- 10 = Society: Demographic trends; the impact of social policies on particular areas of the country or social strata; the incidence of crime or allegations thereof; the impact of media and/or religion.
- 88 = Other

4. Article Length

How long is the article?

- Open-ended: Enter the word count of the article. (Copy and paste the body of the article into <https://wordcounter.net/> and use its count. Be careful not to include captions on pictures, etc.)

5. Number of Data Sources

How many unique data sources are listed in the article/graphics?

- Open-ended: Only code unique data sources that are either (a) listed at the bottom of the article or (b) listed at the bottom of a graphic.

6. Data Source Type(s)

Which of the following types of data sources were explicitly referenced in the story?

(Only code unique data sources that are either (a) listed at the bottom of the article or (b) listed at the bottom of a graphic.) This may require searching Wikipedia or visiting the source's page. Data sources are those individuals, organizations, or documents that serve as the basis for quantified claims or assertions. If a listing of data sources is present in the article (usually at the bottom), only code the listed sources. Otherwise, you may code sources in the story. (Check all that apply.)

- Business/Industry: Privately held companies or corporations, or industry trade groups (e.g., non-profits with a mission statement clearly advocating for an industry or demonstrating representation thereof), such as Uber or the American Ambulance Association. Polling companies (e.g., Rasmussen and Gallup) should be counted as a business source.
- Educational/Academic: Data from academic studies or institutions, such as the University of Massachusetts Amherst or a publication in *Nature*. Data from websites ending in .edu, appearing in .org websites affiliated with a university, or appearing in academic journals should be counted as Academic.
- Government: Data from a governmental body or public bureau, such as local, state, and national agencies, as well as official intergovernmental bodies like Eurostat, the United Nations, or the World Bank (because its members are governments).
- Non-profits: Data from non-profit organizations, including public interest groups (e.g., Greenpeace) and partisan/non-partisan think tanks (e.g., Center for Responsive Politics). You may check if an organization is a non-profit on <https://www.guidestar.org>.

- Self-collected: Data collected by the news organization or journalist.
- Other News Organization: Data collected by another news organization (or attributed to another news organization, even if someone else collected it).
- Other: Data collected from some other source type.

7. Kind(s) of Data

What kind of data were used/presented in this item? (Check all that apply.)

- Geodata: Data that details the location of phenomena/incidents. If a map is included as a visualization, geodata is automatically present.
- Financial: Data about economic phenomena, market movements, government budgets, or financial disclosures, such as stock market values, annual filings, tax returns.
- Metadata: "Data about data," including information about application/service use, the number of followers an individual has, etc.
- Personal: Data about an individual or small number of identified people, such as biographic data, weight and height of sportsmen, and the income of the richest politicians.
- Population Count: Data on the incidence of a particular social phenomenon among a population (e.g., number of border crossings, citizens shot by police officers, etc.).
- Sensor: Data automatically captured by scientific measurement instruments, including aircraft noise, weather data, pollution levels, and train speeds.
- Sociodemographic: Data on features of populations and groups, such as their religious preferences, income, education levels, and employment status.
- Sports Statistics/Play-By-Play: Data on a player's or team's performance across different sports-related metrics (e.g., batting average, win percentage); or modeled on play-by-play data (e.g., to capture 'swings in a game').
- Survey: Data collected about groups' opinions and personal views, compiled through a survey/poll. This includes polls as well as survey data from academic studies.
- Other: Any other kind(s) not listed above (please describe)

8. Geography of Data

What is the geographical reference for the majority of the data analysis? If the source data goes up to the national scope, but the analysis focuses on state-level comparisons, code it as state-level. Give preference to a news organization's self-designation such that stories appearing in the "National" section or with a "national" designation in the URL are coded as National.

- 1 = Global: It covers a large number of countries or all nations.
- 2 = International: It covers a small number of countries, such as the USA and Germany
- 3 = National: It covers a single country, such as Switzerland

- 4 = Regional: It covers a large region (e.g., New England) or a state (e.g., Massachusetts)
- 5 = Local: It covers a city, such as Boston
- 6 = Hyperlocal: It covers a portion of a city or town, such as a neighborhood
- 99 = Not applicable: There is no clear geographical reference in this article (e.g., focuses on Donald Trump's tax returns, but with no reference to a broader impact).

9. Purpose(s) of Data

To what purpose are data used in this item? (Check all that apply.)

- Comparisons: Values are compared across groups, such as Tom Brady's performance vs. Drew Brees'.
- Connections and flows: The strength of relationships and associations are described, such as the likelihood a candidate will follow a particular path to the White House
- Rankings/Hierarchy: Data are recoded to reflect prioritization or described through rankings, such as what respondents believe to be the most important issue government should address.
- Trends: Patterns and trends are described through the use of longitudinal data

10. Appendices

Is additional information provided about the data source(s), data collection method(s), or analysis through an external link, pop-up, or dedicated area on the page? (This may include a link to the organization's GitHub page or a "Methodology" section at the bottom of the page. This should not include additional information created by a third party, such as a methods section from a report by the United Nations.)

- 0 = No: There is no additional material.
- 1 = Yes: There is/are additional material(s).

11. Data Access

Is it possible to download any of the data used in the reporting? This may include direct (one click to download) or indirect (page with a clearly identifiable download/access link or information contained therein) links to a structured dataset or data tables.

- 0 = No: There is no clear avenue for downloading any of the data used in this item. (Code this as "No" if no links are provided, even if you know the data can be downloaded.)
- 1 = Some data: Some of the data can be downloaded from a third-party through links provided in the item.
- 2 = All data: All of the data can be downloaded from a third-party through links provided in the item.

12. Number of Visualizations

How many data visualizations were there in this article? (A visualization becomes distinct when it is set apart from other visualizations in order to communicate something different. For example, two side-by-side maps may constitute a single visualization if they are intended to convey a comparison. Alternatively, a map of police shootings followed by a block of text and a line graph of police shootings would constitute two different visualizations. Videos may count as a visualization but should only be viewed if the cover image features a visualization. Please note that the visualization should communicate quantified information and not just illustrate an idea.)

- Open-ended: Count the number of distinct data visualizations.

13. Visualization Dimensionality

How many quantitative variables are featured in a visualization? (Only include static variables (i.e., those you don't need to hover over to see). Do not count the 'value' as a variable. Do not count geospatial information as an additional variable (e.g., county may be a variable, but the geometry for that county should not count). Check all that apply.)

- Univariate: The visualization relies on data from a single variable. ([example #1](#), [example #2](#), [example #3](#))
- Bivariate: The visualization relies on data from two variables. (example #1, [example #2](#), [example #3](#))
- Multivariate: The visualization relies on data from three or more variables. ([example #1](#), [example #2](#), example #3)

14. Visualization Data Structure(s)

Which of the following data structures were necessary for a visualization? (Check all that apply.)

- Network: Includes data describing how entities are linked to one another, such as how CEOs are linked to one another via mutual boardroom positions. ([example](#))
- Spatial: Includes data that corresponds to some physical area, such as a location on a map or the shape of a county. ([example](#))
- Temporal: Includes data that illustrate change over time. ([example](#))

15. Visualization Interactivity Option(s)

Are there any interactive affordances for the visualizations on the page? (Check all that apply.)

- Filtering/Sorting: The user can include or exclude information by selecting a set of criteria or deselecting categories, or sort information in a defined order (e.g., high to low).
- Personalization: The user can enter data to personalize the visual, such as a ZIP code or date of birth.
- Play/Pause: The user can play or pause an animated visualization (e.g., timelapse).
- Playful interaction: A quiz or game accompanies the visualization, such as asking the user to guess a pattern before showing it.

- Search: The user can search through entries using open-ended fields.
- Zooming/Details: The user can zoom in and out of the chart, or click or hover on data points for additional information.
- Other: Some other interactive affordance is offered (please describe).