KNOWLEDGE CREATION, DISSEMINATION, AND PRESERVATION STUDIES

Locke, Brandon T. and Jason A. Heppler. 2018. Teaching Data Literacy for Civic Engagement: Resources for Data Capture and Organization. *KULA: knowledge creation, dissemination, and preservation studies* 2(1): 23. DOI: https://doi.org/10.5334/kula.23

REFLECTION

Teaching Data Literacy for Civic Engagement: Resources for Data Capture and Organization

Brandon T. Locke[1] and Jason A. Heppler[2]

[1] Michigan State University, US

[2] University of Nebraska, Omaha, US

Corresponding author: Jason A. Heppler (jheppler@unomaha.edu)

Endangered Data Week emerged in the early months of 2017 as an effort to encourage conversations about government-produced, open data and the many factors that can limit its access. The event offers an internationally-coordinated series of events that includes publicizing the availability of datasets, increasing critical engagement with them, encouraging open data policies at all levels of government, and the fostering of data skills through workshops on curation, documentation and discovery, improved access, and preservation. The reflection provides an outline of the curriculum development happening through Endangered Data Week and encourages others to contribute.

Keywords: open data; civic data; civic engagement; curriculum; data literacy; pedagogy; instruction

The 2016 US presidential election set off a wave of activism surrounding government data, particularly in the collection and mirroring of environment and climate change data. While open access to public data was already hampered by unfunded mandates, political and legal challenges, and benign neglect, the election provoked many fears that data running counter to the incoming administration's political agenda could soon go offline. The mass-downloading and mirroring efforts of groups like Environmental Data & Governance Initiative (EDGI) and DataRefuge made this issue a national news story and generated a lot of momentum for action.

Endangered Data Week (endangereddataweek.org) emerged in the early months of 2017 as an effort to encourage conversations about open government-produced data and the many factors that can limit its access. Endangered Data Week offers an internationally coordinated series of events that focus on ways that data may become endangered due to political, technical, and social factors. These events include publicizing the availability of datasets and increasing critical engagement with them, encouraging open data policies at all levels of government, and fostering data skills through workshops on curation, documentation and discovery, improved access, and preservation.

Barriers to Public Use

The benefits of public datasets to cities, states, and nations can increase significantly if broad and diverse publics are able to access and effectively use them. The open publication of data can be a boon for transparency and openness, economic development, and broad civic engagement and advocacy. However, the benefits of machine-readable public information are unevenly distributed to the public. In order to make effective use of these data, one must know the materials are available, have the required hardware and software to access them, and have the digital (and often statistical) literacy to interpret them. There are persistent barriers to accessing and using public information—time, literacy, social status, contextual knowledge—and we must be mindful not to introduce additional barriers along with new forms of electronic distribution.

The tools and technical knowledge required to collect civic data represent significant barriers to access and therefore to openness, government accountability, and potential economic, social, and policy benefits. The creation of datasets often requires scraping information from the web in flat HTML or confusing databases or extracting non-standard fields from poorly documented databases. Data acquired through either method are often irregularly formatted or melded together from multiple sources, requiring indexing, correction, and reorganization. Meaningful research often requires an iterative process of researching the contexts in which the data were created and the data itself to resolve undocumented meaning in the data. Both contexts also require interpretation for specialized and non-specialized audiences.

We can turn to state-level open data policies to better understand the challenges of open data. According to the Sunlight Foundation, forty-one states have open data policies in place.[1] But the open data vary from place to place: some civic institutions provide their data as tables on websites, while others have developed ways to either download machine-readable data directly or access the data through an application programming interface (API). Douglas County, Nebraska, for example, which includes the Omaha metropolitan area, provides an open data portal to some of their data on subjects such as zoning, lead hazard registry, and spatial datasets for streets, boundaries, planning, and elevation.[2] The portal includes APIs as well, which allow users to extract data already collected by the county for use in their own analysis or project, but the data available to citizens can be uneven or difficult to leverage. The City of Omaha's police department provides basic information on offenses, calls, and traffic stops, but the data are provided as a yearly summary in a PDF table rather than an individual breakdown in machine-readable text.[3] In both of these cases, accessing and reformatting the data for reuse requires a fair degree of technical knowledge.

Towards a Data Literacy Curriculum

Another key tenet behind Endangered Data Week is exploring what *can be done* with the data after it has been acquired. The acquisition of public data can have many uses, from identifying patterns of discrimination to looking for instances of missing information or creating visualizations. The barriers to overcoming disorganized data and preparing it for analysis and visualization have been lowered in recent years through both off-the-shelf tools and programming languages designed to manipulate and visualize information. We have documented a variety of these tools and methods through workshops during Endangered Data Week.

A data literacy curriculum must be pragmatic, but not prescriptive. It must be focused on tools and workflows with an understanding that actual usage may vary or other tools and techniques must be employed. Organization, formatting, and correction are fundamental steps in any kind of data usage, but there is no simple and clear process. There is no 'correct' way to organize data—although it is best to follow standard conventions whenever possible, organization should be seen as a means to conduct analysis or make the data more useful for others. This process should be informed by knowledge of the data collection (as much as is possible with public data), the types of questions one wishes to ask of the data, and knowledge of the analytical tools to be used. This is often an iterative process, as new information comes to light or goals shift. Because organization and preparation are all a part of the research process, it is also crucial to document these steps for transparency and reuse.

Oftentimes, these data will come to users quite disorganized: data elements may not be cleanly divided into columns and rows; or the column layout may not reflect the fields that are necessary for manipulation and visualization tools (see **Figure 1**, for example). These messy datasets may not even be formatted as tables made up of columns and rows. There may be issues with the values in datasets as well: inconsistency among dates and/or times, lists of items separated by unique or inconsistent separators (like commas or semicolons), misspellings, incorrect latitude/longitude coordinates, or mixed values (such as integers and characters). Often the first step in working with any particular dataset is to clean the data before beginning analysis.[4]

Data cleaning can often be undertaken with user-friendly tools such as Excel or OpenRefine, while trickier datasets may require a programmatic approach such as using the R programming language for cleaning and manipulating data for analysis and exploration. Whichever tools are used for data manipulation and organization, data will often need to conform to a standard set of guidelines in order to be analyzed and visualized. As the statistician Hadley Wickham suggests, data should be organized according to the following criteria:

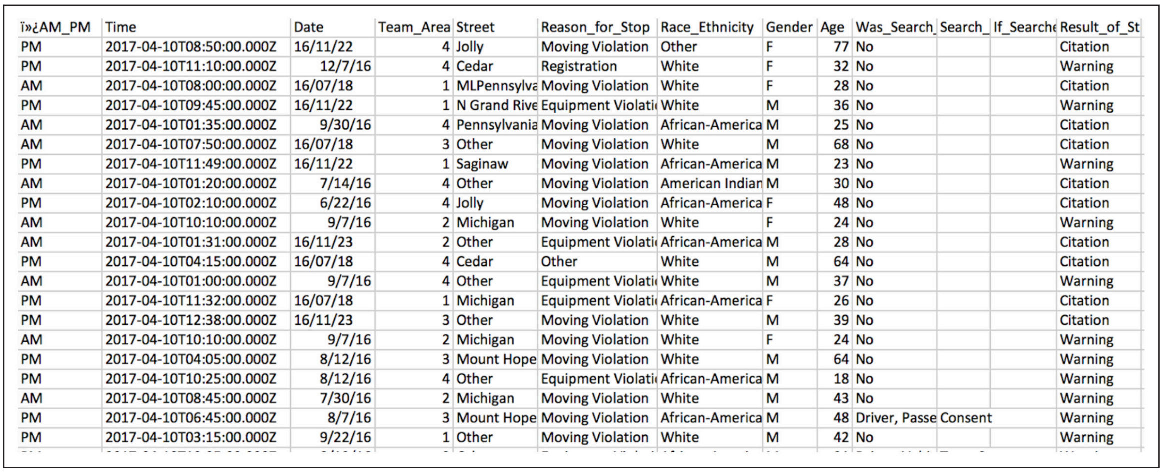
- Column headers need to be variable names rather than values
- Multiple variables need to be separated into their own columns
- Values or observations need to be separated into their own rows
- Different observational units need to be separated into columns or new tables

[1] Open Data Policy Collection, "List of all open-data policies," accessed May 7, 2018, http://www.opendatapolicies.org/browse/all/.

[2] Omaha/Douglas County Geographic Information Systems, accessed May 9, 2018, https://data-dogis.opendata.arcgis.com/.

[3] See, for example, the Omaha Police Department Annual Reports, accessed May 9, 2018, https://police.cityofomaha.org/crime-information/annual-reports.

[4] The concept of "tidy data" and the discussion below on data guidelines is borrowed from Hadley Wickham, "Tidy data," *Journal of Statistical Software* 59 (2014). DOI: 10.18637/jss.v059.i10.



ï»¿AM_PM	Time	Date	Team_Area	Street	Reason_for_Stop	Race_Ethnicity	Gender	Age	Was_Search_	Search_	If_Searche	Result_of_St
PM	2017-04-10T08:50:00.000Z	16/11/22	4	Jolly	Moving Violation	Other	F	77	No			Citation
PM	2017-04-10T11:10:00.000Z	12/7/16	4	Cedar	Registration	White	F	32	No			Warning
AM	2017-04-10T08:00:00.000Z	16/07/18	1	MLPennsylva	Moving Violation	White	F	28	No			Citation
PM	2017-04-10T09:45:00.000Z	16/11/22	1	N Grand Rive	Equipment Violati	White	M	36	No			Warning
AM	2017-04-10T01:35:00.000Z	9/30/16	4	Pennsylvania	Moving Violation	African-America	M	25	No			Citation
AM	2017-04-10T07:50:00.000Z	16/07/18	3	Other	Moving Violation	White	M	68	No			Citation
PM	2017-04-10T11:49:00.000Z	16/11/22	1	Saginaw	Moving Violation	African-America	M	23	No			Warning
AM	2017-04-10T01:20:00.000Z	7/14/16	4	Other	Moving Violation	American Indian	M	30	No			Citation
PM	2017-04-10T02:10:00.000Z	6/22/16	4	Jolly	Moving Violation	African-America	F	48	No			Citation
AM	2017-04-10T10:10:00.000Z	9/7/16	2	Michigan	Moving Violation	White	F	24	No			Warning
AM	2017-04-10T01:31:00.000Z	16/11/23	2	Other	Equipment Violati	African-America	M	28	No			Citation
PM	2017-04-10T04:15:00.000Z	16/07/18	4	Cedar	Other	White	M	64	No			Citation
AM	2017-04-10T01:00:00.000Z	9/7/16	4	Other	Equipment Violati	White	M	37	No			Warning
PM	2017-04-10T11:32:00.000Z	16/07/18	1	Michigan	Equipment Violati	African-America	F	26	No			Citation
PM	2017-04-10T12:38:00.000Z	16/11/23	3	Other	Moving Violation	White	M	39	No			Citation
AM	2017-04-10T10:10:00.000Z	9/7/16	2	Michigan	Moving Violation	White	F	24	No			Warning
PM	2017-04-10T04:05:00.000Z	8/12/16	3	Mount Hope	Moving Violation	White	M	64	No			Warning
PM	2017-04-10T10:25:00.000Z	8/12/16	4	Other	Equipment Violati	African-America	M	18	No			Warning
AM	2017-04-10T08:45:00.000Z	7/30/16	2	Michigan	Moving Violation	White	M	43	No			Warning
PM	2017-04-10T06:45:00.000Z	8/7/16	3	Mount Hope	Moving Violation	African-America	M	48	Driver, Passe	Consent		Warning
PM	2017-04-10T03:15:00.000Z	9/22/16	1	Other	Moving Violation	White	M	42	No			Warning

Figure 1: An example of conflicting and inconsistent data. While the data are organized in rows and columns, the information itself is not particularly standardized (note, for example, the Time and Date columns and the differences between them.) "Traffic Stops." Lansing (MI) Police Department. Accessed March 10, 2018.

The ordering of data into tidy data allows an easier scanning of the information by humans, but also puts the data in a format that can be easily manipulated by computers. Once the data are tidy, analysis and visualization can be easily conducted. As part of Endangered Data Week instruction, we developed a variety of workshops to teach people how to acquire, manipulate, and use data. Some of the workshop material is available in our GitHub repository, such as:

- A Brief Introduction to Federal, State & Local Data: This is a brief slide desk designed to give a high-level idea of the types of public data available, as well as some of the challenges to using them.
- OpenRefine for Complicated Data: This workshop is designed to teach strategies for organizing, geocoding, and normalizing civic data with the open source tool OpenRefine.
- Data Manipulation with R: This hour-and-a-half workshop is designed to introduce people to some basics about the R programming language and the tidyverse packages used for manipulating and tidying data. The workshop includes slides for introducing the concepts of tidy data and data manipulation as well as an interactive R Markdown notebook for working through the concepts.
- Web Scraping with R: Like the data manipulation workshop, web scraping with R includes slides introducing people to the rvest package designed for web scraping, as well as an interactive R Markdown worksheet.

We are continuing to compile new resources, workshops, and other curricular activities at our Resources repository on GitHub and encourage anyone interested in working with data to contribute. The development of this curricula continued through the 2018 Endangered Data Week's events and will continue following our selection as part of the fifth round of Mozilla's Open Leadership mentors program, which is designed to help project leads direct open source projects. After all, our goal is to help people feel empowered about working with data, understand the problems with data, and learn how data can be repressed, lost, or destroyed.

Competing Interests

The authors have no competing interests to declare.

References

Eubanks, Virginia. 2017. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. First Edition. New York, NY: St. Martin's Press.

Forsythe, Diana, and David J. Hess. 2001. *Studying Those Who Study Us: An Anthropologist in the World of Artificial Intelligence*. Stanford, Calif: Stanford University Press.

Gillespie, Tarleton. 2010. "The Politics of 'Platforms.'" *New Media & Society* 12(3): 347–64. DOI: https://doi.org/10.1177/1461444809342738

Gitelman, Lisa. 2006. *Always Already New: Media, History and the Data of Culture*. Cambridge, Mass: MIT Press.

Hicks, Marie. 2017. *Programmed Inequality: How Britain Discarded Women Technologists and Lost Its Edge in Computing*. Cambridge, MA: MIT Press.

Johnson, Eric, and Alicia Kubas. 2018. "Spotlight on Digital Government Information Preservation: Examining the Context, Outcomes, Limitations, and Successes of the DataRefuge Movement." *In The Library With The Lead Pipe*. (Feb. 7, 2018). http://www.inthelibrarywiththeleadpipe.org/2018/information-preservation/. Archived at: https://perma.cc/42LB-VZ68.

Kitchin, Rob. 2014. *The data revolution: Big data, open data, data infrastructures and their consequences*. Los Angeles: Sage.

Noble, Safiya. 2013. "Google Search: Hyper-Visibility as a Means of Rendering Black Women and Girls Invisible." *InVisible Culture: An Electronic Journal for Visual Culture* 19. (October 29, 2013). http://ivc.lib.rochester.edu/google-search-hyper-visibility-as-a-means-of-rendering-black-women-and-girls-invisible/. Archived at: https://perma.cc/9V44-TZU8.

Noble, Safiya. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. First edition. New York: Crown.

Rosenberg, D. 2013. "Data Before the Fact." In: *"Raw Data" Is an Oxymoron*, Gitelman, Lisa (ed.). Cambridge, Massachusetts: The MIT Press.

Tactical Technology Collective. 2016. *Decoding Data*. https://exposingtheinvisible.org/guides/decoding-data. Archived at https://perma.cc/E63U-GWWT.

Wachter-Boettcher, Sara. 2017. *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech*. First edition. New York, NY: W.W. Norton & Company.

Wreyford, Natalie, and Shelley Cobb. 2017. "Data and Responsibility: Toward a Feminist Methodology for Producing Historical Data on Women in the Contemporary UK Film Industry." *Feminist Media Histories* 3(3): 107–32. (July 1, 2017). DOI: https://doi.org/10.1525/fmh.2017.3.3.107

How to cite this article: Locke, Brandon T. and Jason A. Heppler. 2018. Teaching Data Literacy for Civic Engagement: Resources for Data Capture and Organization. *KULA: knowledge creation, dissemination, and preservation studies* 2(1): 23. DOI: https://doi.org/10.5334/kula.23

Submitted: 20 February 2018 **Accepted:** 18 May 2018 **Published:** 29 November 2018

]u[*KULA: knowledge creation, dissemination, and preservation studies* is a peer-reviewed open access journal published by Ubiquity Press

OPEN ACCESS