2 comments

Putting data back into context

Why context is hard

04 April 2019

By Catherine D'Ignazio

What happens when an institution collects data about something in the world? The origin of the word *data* actually means 'that which is given'. And this is typically how newcomers regard data – as a somewhat neutral recording of facts at hand. The information was there, and then an institution collected and stored it. When data journalists investigate an issue, we look for who might have data, how we can acquire those data, and then use them to create new insights into the world.

But the scholar Johanna Drucker proposes a different word for data: *capta*. By this, she means, 'that which is taken'. As Johanna states in her paper, Graphesis: Visual knowledge production and representation, "Data are

experiment".

This distinction might seem academic for data journalists, but in fact it's at the root of why *context* matters so deeply for data journalism. Thinking of data as *capta* invites us to consider why an institution invested their resources in collecting information, how the institution uses that information, who the information benefits (and who it doesn't), and what the potential limitations of the information are. In short, it points us back to how data are never neutral 'givens', but always situated in a particular context, collected for a particular reason. In Lauren Klein and I's book, called Data Feminism, we devote an entire chapter to the importance of considering context, particularly when the collection environment has any kind of power imbalances.

Biography

Catherine D'Ignazio is an Assistant Professor of Data Visualization & Civic Media at Emerson College and a research affiliate at the MIT Center for Civic Media and the MIT Media Lab.

Why context is hard

Establishing and understanding the context of your data (capta) is likely one of the single most challenging aspects of doing data journalism. It's like starting out with the leaves of a tree and then trying to connect them back to their branches and roots. But why is context so hard?

First of all, data are typically collected by institutions for internal purposes and they're not intended to be used by others. As veteran data reporter Tim Henderson, quoting Drew Sullivan, said to the NICAR community, "Data

institution, not from the perspective of a journalist looking for a story. For example, one semester my students spent several weeks trying to figure out the difference between the columns 'PROD.*WASTE*(8.1_THRU_8.7)' and '8.8_ONE-TIME_RELEASE' in a dataset tracking the release of toxic chemicals into to the environment by certain corporations. This is not an uncommon occurrence!

And while the open data movement has led to governments launching more open data portals and APIs, these efforts have prioritised publishing data over publishing the metadata that would actually make the data more useful to outsiders. Part of this is cost-related -- context is expensive. The cities of Boston and Chicago both had to secure external grants from foundations in order to embark on comprehensive metadata projects to annotate the columns of the open datasets on their portals and make their datasets easier to search and find.

But sometimes, the lack of attention to usability, context and metadata works in favour of the collecting institution, which may have reasons for why it doesn't want certain information to become public. For example, the Boston Police Department (BPD) runs a programme called Field Interrogation and Observation (FIO). For all intents and purposes, this is a stop and frisk programme, in which police log their encounters –-- observations, stops, frisks, interrogations -- with private individuals on the streets. In 2014, following a lawsuit won by the American Civil Liberties Union, the BPD was obligated to release their FIO data publicly on Boston's data portal. But when you search for 'stop frisk' on the portal, nothing comes up. Journalists and members of the public would need to know the bureaucratic term for the programme (FIO) in order to be able to locate it on the portal.

certain data may be harder to find

Furthermore, some institutions may publish their data and metadata freely, but be less forthcoming about their data's limitations. This can lead to serious cases of misuse and misrepresentation of data. In one chapter of Data Feminism, The Numbers Don't Speak for Themselves, Lauren Klein and I discuss the case of GDELT: the Global Database for Events, Language and Tone. In a high-profile correction, FiveThirtyEight had to retract a story about the kidnappings of Nigerian girls that used the GDELT database. They had mistakenly used media reports about kidnappings to tell a story about the incidence of kidnappings. While FiveThirtyEight should have verified their data before publishing, we describe how GDELT, because of their pressure to attract big data scientific research funding, failed to describe the limitations of their 'events' data (which is not events data at all, but rather 'media reports about events' data).

The Three-Step Context Detective

So, what's a data journalist to do? She has to become a 'context detective', working with data that have been captured from the world into spreadsheets and databases, and connecting them back into their collection environment. This work is similar to that of a detective -- the journalist has to use incomplete clues that point backwards to the bureaucratic functionings of the collecting institution.

To understand the data, you must understand the who, what, when, where, why, and how of that bureaucracy. Below I present a process called the Three-Step Context Detective, which I use in the classroom. These steps don't necessarily have to be completed in this order.

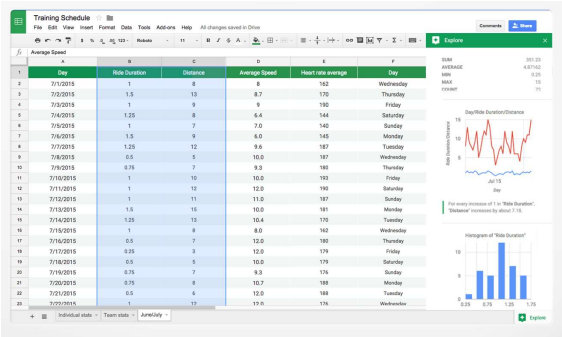
1. Download the data and get orientated

Looking at hundreds, thousands or hundreds of thousands of obliquely named rows and columns can be daunting at first. Sometimes newcomers think that the data science 'wizards' can just look at a spreadsheet and see

can ask good questions.

You can use a programme like Excel or Google Sheets to do basic exploration to answer questions like:

- How many observations (rows of data) do you have?
- How many fields (columns) do you have?
- Is it clear what each row is counting? (Remember those incidences of kidnapping versus media reports about kidnapping – getting crystal clear about what your data is logging is supremely important.)
- What is the time period of the data? Use the 'Sort' function on any column with dates or timestamps to see when the data begin and when they end.
- What is the geographic extent of the data?
- Does there appear to be a lot of missing data?



Additional resources

For more on using spreadsheets, check out Brant Houston's article Spreadsheets for journalism, or our video courses Doing Journalism with Data: First Steps, Skills and Tools and Cleaning Data in Excel. You can also start a conversation in our forums.

This process of getting oriented with a new dataset is no small task. In fact, Rahul Bhargava and I built a free, online tool called WTFcsv which captures the emotion that journalists often feel when looking at a new spreadsheet: "WTF is going on with my csv file?!" Using WTFcsv, you can continue your orientation process.

image below depicts data about passengers on the Titanic. The column ‘Sex’ is rendered as a column chart that demonstrates, visually, that there were 314 female and 577 male passengers logged on the Titanic.

Rahul and I talk about the importance of asking good questions of your data before you conduct analysis or tell stories. WTFcsv can help you answer all of the basic questions above, as well as start to form your own questions about the dataset in front of you. For example, for the Titanic data, good questions might be about ethics (‘Why is 'Sex' a binary variable?’), data formatting (‘What does the 'Parch' column mean?’), data quality (‘Is this data complete?’), or data analysis (‘Did women survive at a higher rate than men?’).

It's important to write down all of these questions, because as you go through the next couple steps, you can try to answer them.

2. Explore all available metadata

Metadata is data about data and can be your golden ticket to establishing context for a dataset. In an ideal world, any dataset that you download would have a detailed and up-to-date data dictionary. This is a document that provides a column-by-column description of the dataset, along with guidelines for using it.

ShallowFlag	3	meters. A value of 0 denotes a coral that has a correct DepthInMeters greater than or equal to 50 meters.	1; 0; -999	This is created to capture deep water taxa that reach into shallower areas.	0
DatabaseVersion	4	Version of the entire database indicated as a date-based version in the format YYYYMMDD_<iteration>. Example: '20190226_0' for the version created on February 26, 2019. The zero on the end indicates the iteration number on that day. If another version of the database was created on that same day, it would be indicated as '20190226_1'.	[value]	In ERDDAP as a global variable	0
DatasetID	5	Standardized ID for dataset. The form of this ID is specfied in a separate internal document.	[value]	This value can be thought of as a standardized version of a collection level ID for the data set. In some cases we break the data down below the collection level if the collection is very large and has differing methods of sampling (see SamplingEquipment term)	0
CatalogNumber	6	Unique record identifier assigned by the DSCRTP. It is persistent and the numbers are retired if records are deleted from the database.	[value]; -999	NA	0

In the above example, from the NOAA National Database of Deep-Sea Corals and Sponges, each field (column) in the dataset is annotated and explained, along with descriptions of data quality, units of measurement, completeness, and usage guidelines.

Seeking metadata is not always easy or successful. Not all data providers produce data dictionaries. In 2016, journalist J. Albert Bowden sought documentation on the fields in a dataset from the US Department of Agriculture. He was told that explanation of their column headers was a proprietary secret. Moreover, even when there are metadata, providers might fail to call the file 'data dictionary'. For example, if you use the City of Boston's 311 data, the data dictionary is called 'CRM Value Codex' – not the most attractive and user-friendly name ever.

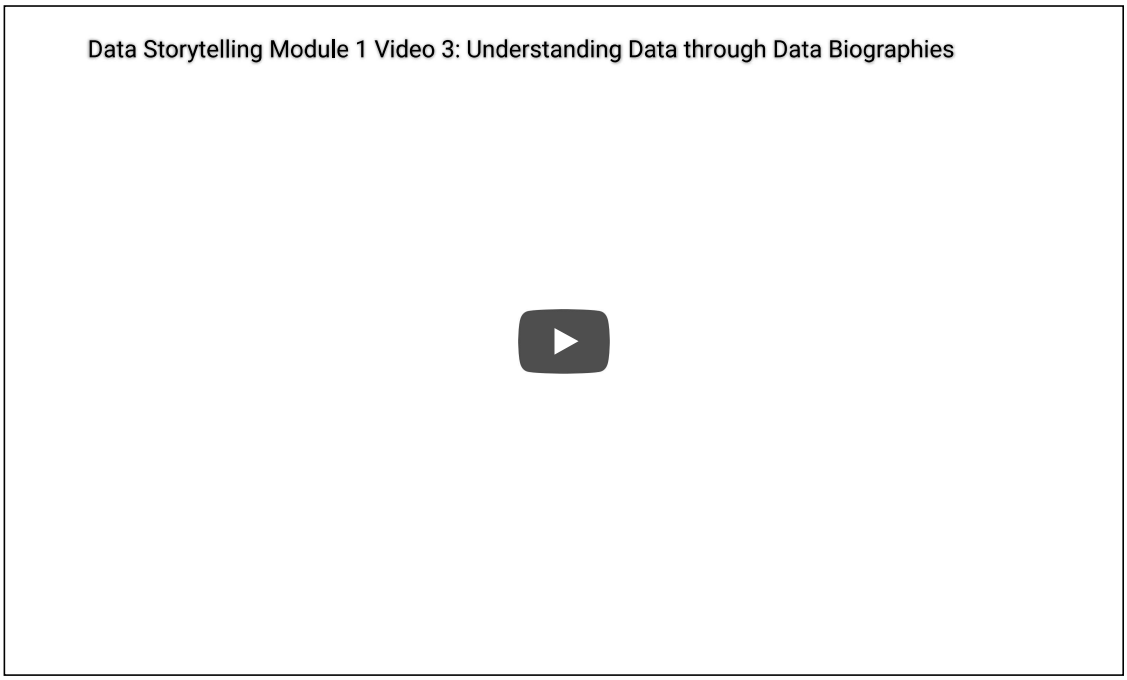
And sometimes data dictionaries or other forms of metadata might be outdated because the institution fails to update them when the dataset changes. It's important to have your sceptical, fact-checking, data-verifying journalist hat on at all times.

3. Background the dataset

Journalists often 'background' a person or 'background' an issue, and likewise the final step of the Three-Step Context Detective is to background your dataset. This may be the most time-consuming aspect of establishing context for your data, but it is well worth the investment in terms of understanding limitations, preventing errors, and discovering newsworthy stories and analysis. In this process, there are at least three separate things to conduct background research on.

Background the collection process

advocates for creating data biographies that describe where the data came from, who collected them, and how they collected them. In the case of the Boston stop and frisk programme discussed above, police officers fill out paper forms after having an encounter with a resident on the street. Then, those forms get turned into the precinct and a staff member logs the values in a database. Before publishing to the website, other staff members remove personally identifying information. It's all very mundane, but absolutely essential to understanding where errors and missing data can be introduced. The meat is in the bureaucratic details like whether data is self-reported, or observed, or measured by a machine. Like what database the organisation uses to store the data. Like whether the way the organisation is counting and measuring has changed recently (making current data and prior data not comparable).



Heather Krause describes how to create a data biography, and why it can be essential to understand what the data do, and do not, actually measure.

After exploring any available metadata, your best path to backgrounding the collection process is finding a human being to talk to. This might be someone from the collecting organisation, but it's important to think creatively about interviewees when this is not feasible.

someone from the collecting organisation, but it's important to think creatively about interviewees when this is not feasible. For example, in the case of stop and frisk data, it's hard to get police spokespeople on the phone. Other potential interviewees might include the youth of colour who are disproportionately stopped, the ACLU who sued the police department and did their own background research on the collection process, or criminal justice scholars who have studied the Boston programme. Krause has a helpful data biography template that you can fill out for this process.

Background the organisation

While we have talked a lot about understanding *datasets*, Yanni Loukissas makes the case in his book, All Data Are Local, that to use data effectively we also need to understand *data settings*. Backgrounding the dataset is not just about the data itself – it's also about understanding why an organisation was motivated to collect it in the first place, as well as how they use it.

In the case of the stop and frisk data, this means doing background research on the Boston Police Department: What is their mission? How long they have existed? What is their budget? How many officers are there? When have they been in the news in the last ten years and why? It also means researching the FIO programme specifically: When and why did the BPD start the programme? Was it part of a national wave of FIO programmes? Is there scholarly and legal debate on whether these programmes are constitutional and effective in reducing crime?

From this understanding of the underpinning organisational and programmatic goals, it's helpful to try to understand how the organisation uses the data it collects internally. For example, does the BPD use its logs of police-civilian encounters to try to limit racial profiling? Do officers have quotas? Who does the BPD have to report their numbers to?

Here, again, interviews with real, live human beings are going to be one of the most effective ways of getting information about the organisation's motivations and uses of the data it collects. But when you can't get an inside interview, one of the best ways to find this information is to:

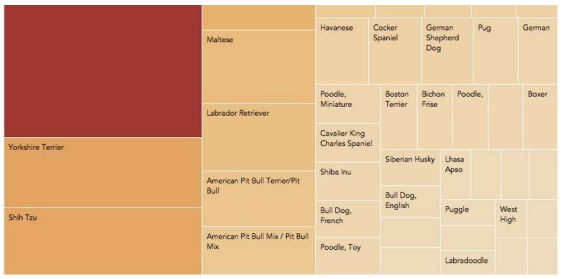
Data is expensive to collect, maintain, and organise. Most organisations don't collect data because they want to but rather because they have to, either to comply with laws or with internal policies. Doing background research on the regulatory environment can often shed light on why the organisation collects data, who it reports that data to, and how it reports the data. For example, all accredited higher education institutions in the US have to collect and report data about sexual assault on college campuses because of the Jeanne Clery Disclosure of Campus Security Policy and the Campus Crime Statistics Act (Clery Act).

It is typically easier to background federal and state laws, which tend to be available publicly or identified via talking with lawyers and others with legal knowledge. Internal policy documents that guide data collection can be harder, albeit not impossible to access. If you live in a country with public records laws, using those laws to request organisational governance documents and training manuals can be an excellent way to understand the internal regulatory context that guides information collection. As an example, most police departments in the US collect data on the use of force by police officers against civilians. Knowing this, when a white male police officer used excessive force at a pool party in 2015, reporters at MuckRock made a public records request for the McKinney police use of force policy. On page nine, the policy details when and why officers are required to file a 'Response to Resistance' report (RTR) and who is responsible for maintaining those reports. This policy would be essential background information for any journalist seeking to write a data story about use of force from RTR data.

Pitfalls

So, the Three-Step Context Detective consists of getting oriented, exploring all metadata and backgrounding the dataset (including the collection, the organisation, and the regulatory environment). In the process of building out these connections between the dataset and its broader context, there are two pitfalls to keep your eyes on.

First, beware of your own brain and its penchant for making



There are many 'Unknowns' in the 2016 dog licensing data from the City of New York. We need to be careful not to make assumptions like 'Unknown' means 'Mixed breed'.

in a dataset to imagine that you know what they mean, but this can be dangerous. In my data journalism class, we were working with data about the dogs of New York City -- their breeds, ages, and sex.

Of course, some fields had incomplete data and 'breed' was one of those with many 'UNKNOWN' values in the column. One student assumed that breed = UNKNOWN meant that the dog was mixed breed and built their whole story around that incorrect assumption ('UNKNOWN' means the information wasn't filled out by the applicant so we literally do not know the breed). Luckily, the student did end up checking their assumptions and revising the story, and the data itself was fairly low stakes. That said, this illustrates the importance of Jonathan Stray's advice about 'considering multiple explanations for the same data, rather than just accepting the first explanation that makes sense'. The same advice applies when assembling the context around your data just as much as it applies when analysing it.

Secondly, it's important to remember that power is not equally distributed across the collection process, the organisation, and the regulatory environment. The result of social inequalities in the data setting is that the numbers may appear to tell one story on first exploration, but that story might be completely false because the collection environment has systematically silenced people with less power. What does this mean?

In The Numbers Don't Speak for Themselves, Lauren Klein and I discuss a story written by three of my students about sexual assault data provided by the Clery Act. What the students found is that campuses with high numbers of sexual assault were not hotbeds of rape culture, instead these campuses were actually providing the most resources and the most supportive environment for survivors to come forward. So, paradoxically, the campuses with the lowest rates of sexual assault were *not* doing great but rather creating an environment that actively discouraged survivors to report. Meanwhile, those campuses with higher numbers were actually

racism, or classism are at work (read: all the time) that lead to the systematic undercounting or overcounting of women and other marginalised groups. The way to address it is through establishing context – the students discovered this through background research, reviewing policy docs and many interviews – rather than accepting the numbers at face value.

Opportunities

Just as there are pitfalls for context, there are also opportunities for journalists and news organisations to create useful resources for readers and other journalists from their work on context. And context is work! Instead of writing a single story from a data exploration, organisations like ProPublica have started to create what Scott Klein calls 'news apps', that is, evergreen resources like Dollars for Docs. While it is useful for individual readers, Dollars for Docs has also become a data resource for other news organisations to write their own, localised stories on the influence of pharmaceutical companies – for example, this story about the effect of pharma money on doctors in St. Louis, Missouri. In this sense, ProPublica has become known as an 'information intermediary', by turning their original investigation's context and data into a resource that is reusable for other organisations.

ProPublica turns the context work that they do compiling and backgrounding datasets into a source of revenue in their data store.

Verified data and expert contextual information can also become a source of revenue for news organisations. ProPublica maintains a data store where you can purchase datasets on a variety of topics. Many of the datasets available come with excellent 'data user guides' – a term coined by Bob Gradeck, manager of the Western Pennsylvania Regional Data Center. In his work promoting open data, he saw the need for metadata that goes beyond the data dictionary to provide a narrative account of a dataset - where it comes from, how it is used by the organisation, and what its limitations are. Examples of Bob's work can be seen in the data user guides for 311 data in Pittsburgh.

The Associated Press (AP) is also getting into the *data + context = revenue* game. They spent extensive time compiling a national database on school segregation in the US, which comes with a 20-page data user guide including where the data is collected from and what kinds of questions it can be used to answer. It's available for purchase, and the AP is starting to develop a subscription model where organisations can pay for access to other datasets, context, and discussions with reporters who worked on those issues.

Conclusion

the story right, even in the face of meager metadata, bureaucratic obstacles, and power imbalances in the data setting. Do you have stories about your work with context and data? Share them in our forums.

Time to have your say

Duncan Anderson · 1 year ago

I've gained so much insight as a data-journalism enthusiast after reading this article. I'm really looking forward to trying out WTFcsv in the near future. Thanks a ton, Catherine!

· Reply

Russell Webster · 1 year ago

Very helpful, indeed Catherine. I'm a specialist (drugs & crime) rather than a journalist so I'm more likely to be sensitive to some of the likely paradoxes like the reporting of sexual assaults you mention. Even so, I've been caught out many times. The most satisfying (albeit rare) component of being a data investigator is when you reveal two parallel data sources, one for public consumption, one the real facts. It is understanding the context that enables you to smell a rat and dig deeper.

· Reply

SIGN IN TO COMMENT

About us
The team
Partnerships
Branding
Contact

Write for us
Contributors
Partners

Latest discussions
FAQ
Newsletters archive

Small print

Terms and conditions
Privacy policy
Code of conduct

Social media

Twitter
LinkedIn
Facebook

Created by European Journalism Centre

Supported by Google News Initiative