RESPONSIBLE DATAFORUM

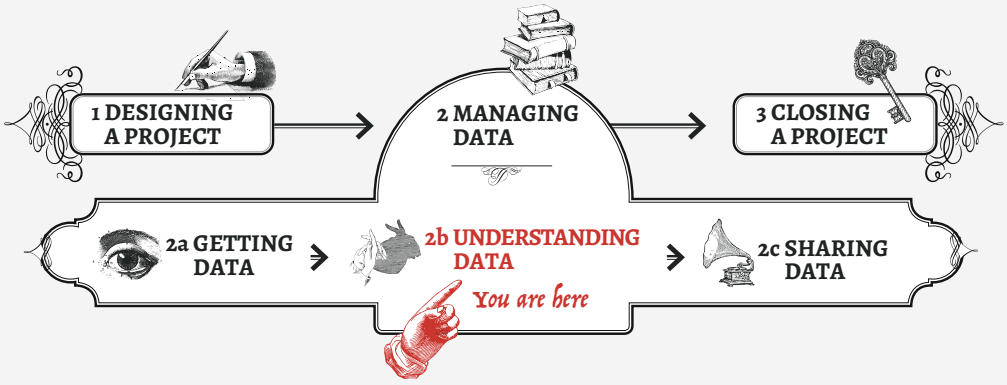
The HAND-BOOK
OF THE MODERN DEVELOPMENT SPECIALIST

26

UNDERSTANDING DATA

CLEANING, PREPARING AND VERIFYING DATA

The project data flow



Target audience

Those who will be working directly with the data, perhaps those with a data science or computer science background, to help them think through the ethical issues that should be considered.

When might this chapter be useful?

Once the data has been collected, as cleaning, preparing and verifying data is discussed in this section.

Content summary

Verifying and cleaning data

Real life data collection is messy. This means that using it can be hard. Ensuring the data you have collected is properly verified is vital. Without verification, the results cannot be relied upon for decision-making processes, and you risk misrepresenting or doing harm to the populations you aim to help.

Some tips around how to *verify* your data are offered; whether you are close to the source, or whether you have no access at all to the original source of data. In both cases, ensuring the *integrity* of the data can help make sure it is interpreted in a responsible way.

Next, *cleaning data* is discussed. Messy data may hide harmful information. If we don't make sure that we can clearly name, describe and recognise all the information contained in data sets, we might make improper assumptions about the risks that that data might pose. This might mean anything from making sure dates are in the same format in a spreadsheet, to ensuring that appropriate filters are applied, or simply organising it in a useful way for what we're trying to find out.

Documenting what you're doing, and preparing data for future use, is then discussed. Collecting appropriate data about the data itself, otherwise known as *metadata* can help others use the data, or simply make sure that others in your team or organisation understand what the data represents. There are also a few specific *standard file types* that will make it easier for your data to be interoperable with other datasets, and used within common applications.

MANAGING BIAS AND ASSUMPTIONS

At this point, the data you're working with should be ready for analysis; you've collected or gathered it, cleaned it, and prepared it for further use. But before going straight on to analysis, this is also a good point to stop and question your assumptions.

All data, no matter how it is collected, will contain a certain amount of *bias* due to, for example, the number of considerations that have gone into collecting the data, the way in which the data is structured, the questions that have been asked, or any number of other considerations.

Specific points that are useful to consider for spotting potential bias are explored here, such as thinking about the *sample* of people from whom data was collected, or cultural biases within the way the collection process was structured. It's good to be on the look out here for any *outliers* within your dataset that might skew your result - that is, datapoints that might represent human or machine errors. Getting expert opinions from those with deeper topical or cultural expertise of where you're working is a good way to verify your findings before moving on, too.

Useful resources

The Verification Handbook *http://verificationhandbook.com*

The Citizen Evidence Lab *http://citizenevidence.org/*

"A gentle introduction to cleaning data"
http://schoolofdata.org/handbook/courses/data-cleaning/

"A gentle introduction to exploring and understanding your data"
http://schoolofdata.org/handbook/courses/gentle-introduction-exploring-and-understanding-data/

The IATI Registry *http://www.iatiregistry.org*

the engine room

This publication is part of a series found at *https://responsibledata.io*, produced by the engine room Responsible Data Program, 2016.