RESPONSIBLE DATAFORUM

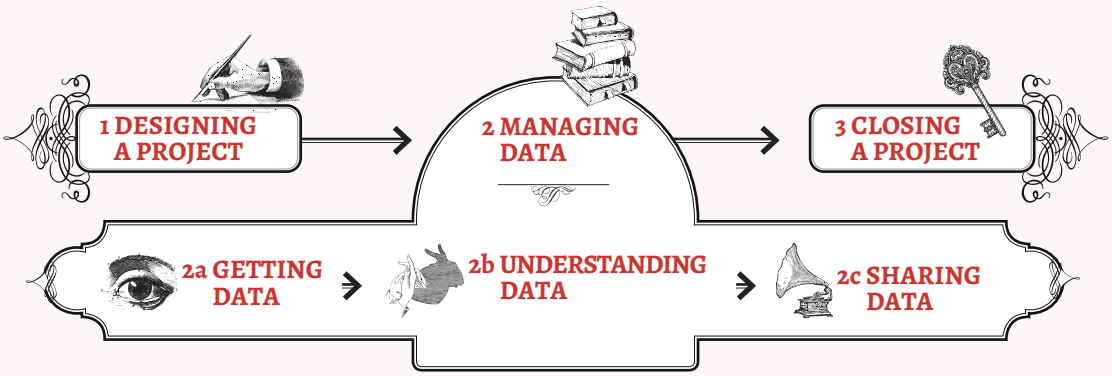
The HAND-BOOK
OF THE MODERN DEVELOPMENT SPECIALIST

Ø

INTRODUCTION

WHAT THIS HANDBOOK HAS TO OFFER.

The project data flow



Target audience

A primer on why *responsible data* is a relevant concern for international development work.

When might this chapter be useful?

If you're wondering what materials this book offers or if you're looking for advocacy material as to why your colleagues or managers should be interested in responsible data.

Content summary

Am I Working with Data?

In short, yes, you are. Data is everywhere, whether we recognise it or not, and it can be qualitative (eg. words or photos) or quantitative (eg. spreadsheets).

No matter the kind of data we're talking about, it's important to remember *the people reflected in the data*. Gathering more and more data on people, even on large groups, can be dangerous. Unknowingly supporting discrimination that might be present in the dataset or, with enough aggregated datasets about the same population, could reveal details about individuals that violates their privacy.

The responsibility to mitigate against these kinds of decisions lies with all of us, not simply with the person who is directly responsible for analysing or collecting the data.

The world of data: opportunities and risks

Data is used in more than just "*data-driven projects*" or "*data journalism*". owadays, all sorts of international development projects rely on data in various formats.hough the rate of data use has increased rapidly over the past couple of decades, the discussion around ethics and politics of data has not evolved at quite the same rate.

Within the development sector, we're seeing increased mention of the "*data revolution*" - but what does this really mean? To date, there's not been so much of a critical debate around the potential harms, or risks, associated with increased data use.

Having data is synonymous with having a certain amount of ***power***. Within development, power disparities between donors and beneficiaries are already well-pronounced, and data could exacerbate those tensions. The data collector - in many cases, the donor - effectively owns a commodity relating to an individual, and that data could be financially valuable. This kind of model is rife within social media networks where the "*if you're not paying, you're the product*" mantra is relatively commonly understood. Similarly, we generally understand that if malicious actors have access to data, it can be dangerous.

This section discusses the role that data can play in either exacerbating, or (hopefully!) weakening power disparities that are present within the development sector, looking at issues like agency, legitimacy and representation. It goes on to discuss how a framing of ***responsible data*** can help, and why we think it is important. At its base, discussions around responsible data centre around ***empowerment*** and ***avoidance of harm***; more than discussions about privacy and digital security, we understand responsible data to be about prioritising dignity, respect and privacy of the people we work with.

To put this in perspective, the chapter touches quickly upon a few of the harms that might come with irresponsible data practice, and goes on to explain a few strategies and arguments as to why responsible data is important. Common tensions that are often misunderstood are also addressed, with appropriate explanations - such as transparency and privacy, or representation and responsibility, to help the reader advocate for responsible data to be prioritised within their organisation.

Often, these advocacy strategies can be met with misconceptions around what responsibility means in this sense, and the very real harm that can be done with misuse of data. A few of the most common misconceptions are addressed here, such as misunderstandings around what ***anonymisation*** actually means in practice; around what data constitutes ***sensitive*** data; and different understandings of what ***privacy*** really means.

Useful resources

The "Am I working with data" flowchart is a fun way to help almost anybody realise that "working with data" is not only for the more technical among us, but can also include words, photos, books - all sorts of things that almost all of us interact with on a daily basis.

Donor code of conduct
http://www.ssireview.org/blog/entry/a_new_donor_code_of_conduct

UN data collection
http://www.unglobalpulse.org/privacy-and-data-protection

Professional standards for protection work carried out by humanitarian and human rights actors in armed conflict and other situations of violence
http://www.icrc.org/eng/resources/documents/publication/p0999.htm

Fair Information Practice Principles
http://www.nist.gov/nstic/NSTIC-FIPPs.pdf

International Principles on the Application of Human Rights to Communications Surveillance
https://en.necessaryandproportionate.org/

OECD Privacy Principles
http://oecdprivacy.org/

UNFPA guidelines on data issues in Humanitarian Crisis situations
https://www.unfpa.org/public/home/publications/pid/6253

the engine room

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

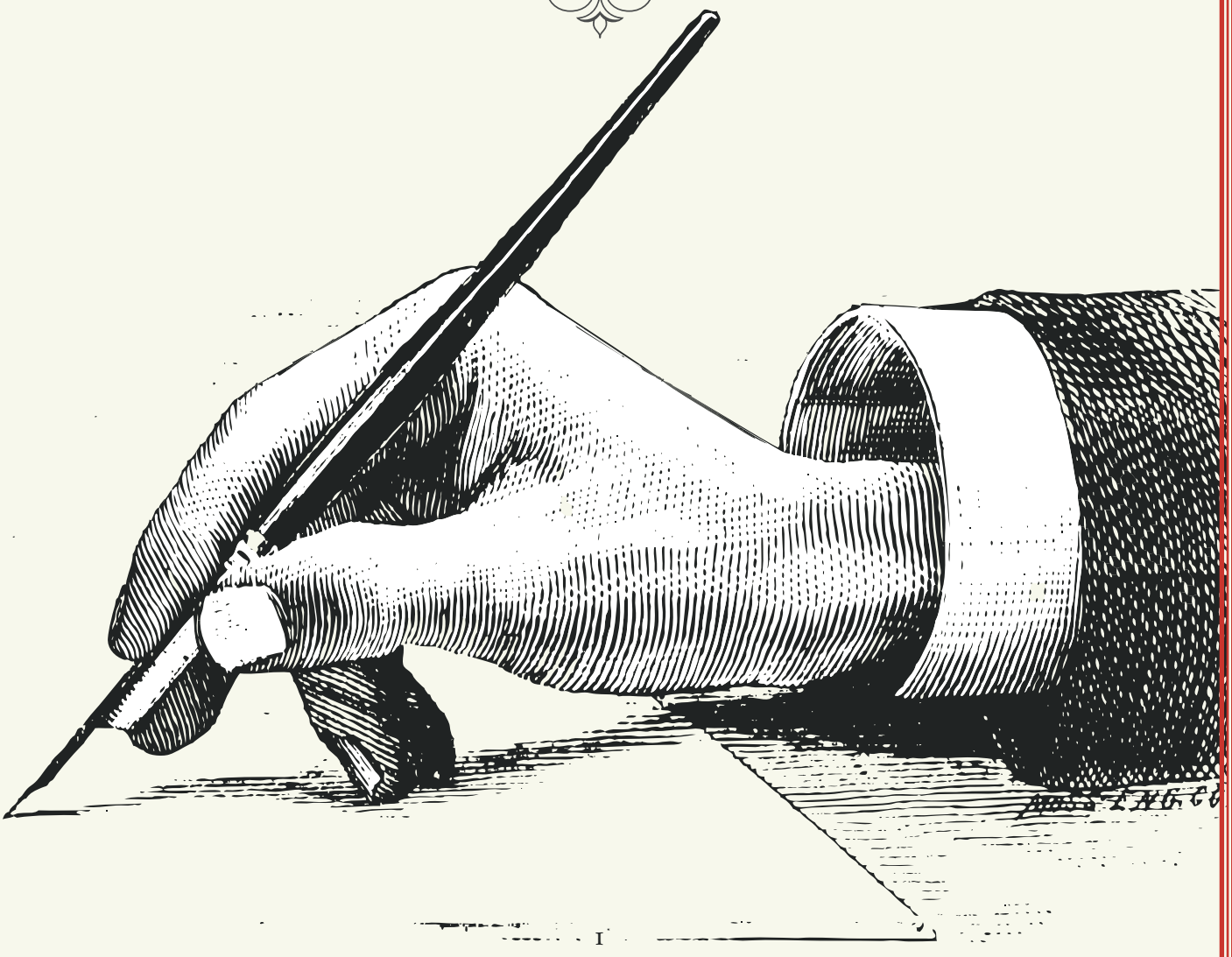
RESPONSIBLE DATAFORUM

The HAND-BOOK
OF THE MODERN DEVELOPMENT SPECIALIST

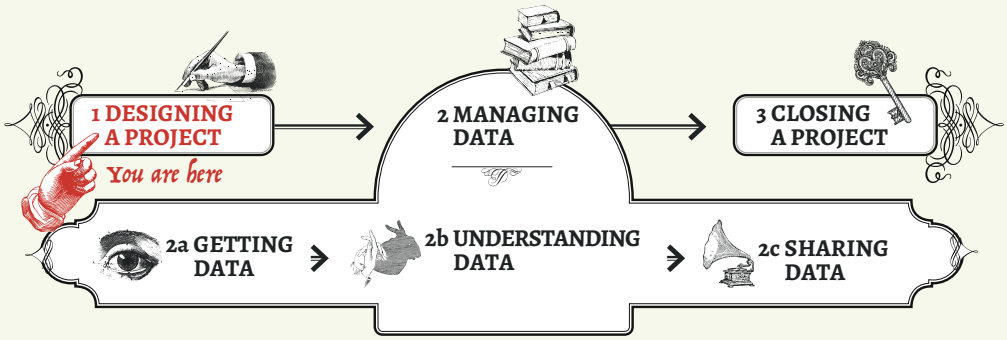
1

DESIGNING A PROJECT

DESIGNING RESPONSIBLE PROJECTS



The project data flow



Target audience

Project managers, or those involved in project design. The second section, "Budgeting" is useful for those who are responsible for drawing up budgets, or for donors who are responsible for looking over and approving project budgets. It contains a number of important responsible data issues that should be addressed in a comprehensive project budget.

When might this chapter be useful?

Towards the beginning of a project; it contains tips and tools to help you embed responsible data practices into the project from the very beginning.

Content summary

The chapter starts with a series of top tips on how to incorporate *responsible data* practices intoa project from the very beginning. It is better to channel energy into comprehensive project planning rather than letting the potentials of what *could* go wrong lead you to panic.

A brief overview of how to conduct a threat modelling exercise, a type of risk assessment, is included here, with main discussion points and potential outcomes. A matrix is suggested as a practical tool to help readers actually carry out this exercise including considerations to bear in mind on how the results of this exercise can be used in practice.

Even with *prevention planning* as discussed above, there are still potential risks that might take place. In case of such incidents, knowing how to discuss them and what to do about them is very important. *contingency planning* is thus also discussed, with broader organisation-wide guidelines to keep in mind when starting a project.

Often, budget constraints are one of the major reasons that responsible data is pushed to the sidelines. This can be countered with a well-thought out budget from the very beginning. Oft-forgotten items are described in this section, such as the responsible data concerns that come with relying upon external providers for tech support; thinking through hosting and storage in a way that is most appropriate to the data that you are collecting; skills required in a team; back up planning and emergency support, and crucially, project termination.

Ending a project well is one of the most often forgotten concerns around a project that might have collected or used a lot of data.This could mean actually archiving the data, making sure that those reflected in the data can have continued access to it, or making sure it is securely taken offline if necessary.

Useful resources

Holistic Security:
Tactical Tech's guide,
https://tacticaltech.org/holistic-security

the engine room

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

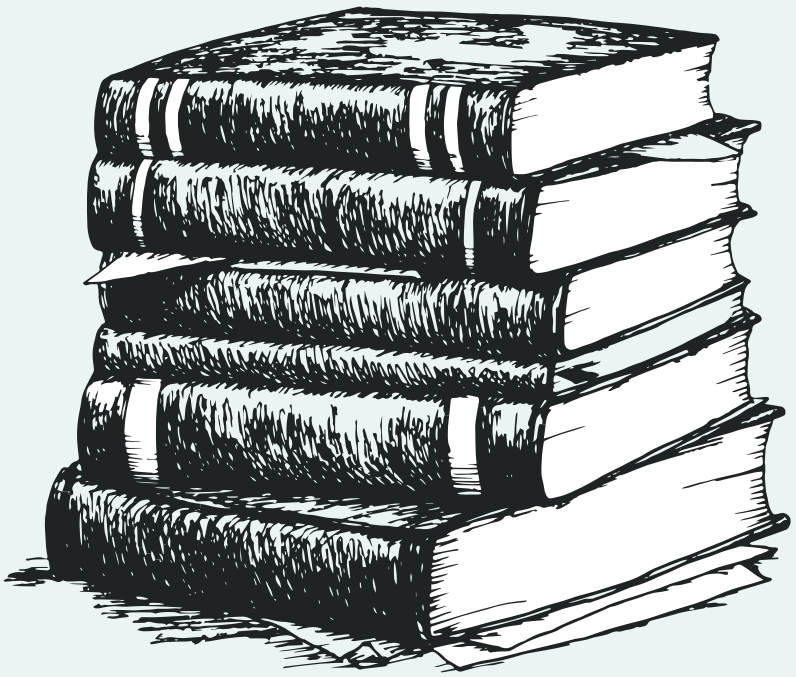
RESPONSIBLE DATAFORUM

The HAND-BOOK
OF THE MODERN DEVELOPMENT SPECIALIST

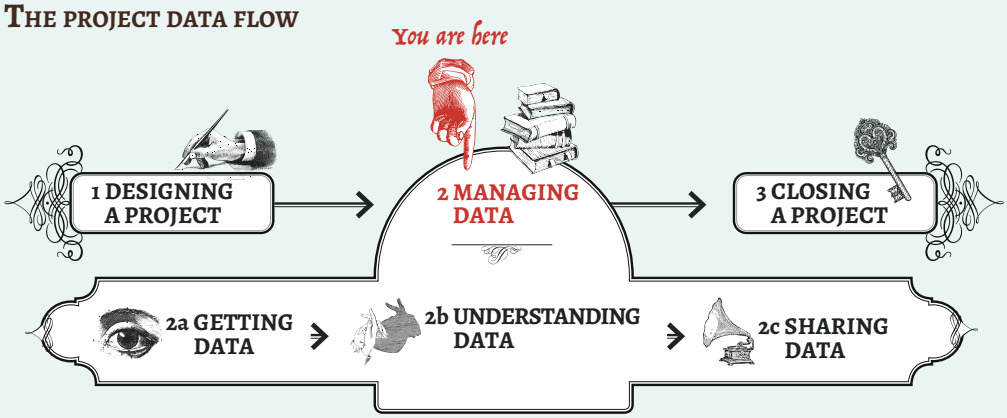
2

MANAGING DATA

SETTING UP THE 'DATA INFRASTRUCTURE'



The project data flow



Target audience

Those who are making decisions, or advising others, on how to go about managing data; from the very basics of deciding where and how to store it, to who has access to it, how it is managed on an ongoing basis, and broad legal considerations to bear in mind.

When might this chapter be useful?

In advance of actually collecting any data, the chapter contains information that might be useful to consider when setting up the 'data infrastructure'.

Content summary

This section provides an overview of approaches you can implement to ensure your information is stored, managed and accessed in a responsible, secure and protected manner. To begin with, it's worth reviewing the general principles of data integrity, and thinking about how to manage risks surrounding data storage.

A home for healthy data

All data occupies physical space, even if we don't think of it as such.

There are lots of decisions and processes that go into creating, storing and sharing data. Some of these are discussed under the section 'data integrity', which looks at the validity, authenticity and security of data. As a frame of analysis to understand what the given data is, and isn't, it can be very useful. Some main considerations to bear in mind when interrogating the *integrity of your data* are included here.

Other risks and harms associated specifically with *data storage*, along with appropriate mitigation strategies, are then discussed. Then, extra considerations to take into account when dealing with *sensitive data* are briefly mentioned.

The pros and cons of various responsible data storage options are then discussed, such as storing data locally, in the cloud, or within a network. Another oft-overlooked aspect of data storage, *physical data storage* is then explored in more detail.Threats to your data don't only happen online. To plan for a possible physical break into your office or headquarters, a checklist is provided of things to consider in advance.

For your eyes only looks at who has access to the data. Essentially, making sure that access to sensitive information happens on a "need-to-know" basis can reduce the risk of someone getting access to data they shouldn't. This can happen online, through setting appropriate user permissions for a certain person's role, or physically, in terms of only letting trusted or vetted individuals into an office space. Regular audits that check who has access and revisiting user permissions can help make sure these access levels remain up to date.

Many of us work often on *collaborative projects*, which require different parties having access to the data. Another checklist to consider when setting up access permissions and data infrastructure on collaborative projects is included here, like including *layered access*, or embedding secure practices into the various access points.

Despite all this, it is important not to over do the checks and security measures in place; when it comes down to it, the project and the data need to be accessible in times of need, and potentially on a longer-term basis. Some concrete steps for making sure that the data will be available for the right people at the right time and planning for disruption or technical emergencies, are then included.

In the final section within this chapter, *Legal Considerations* are discussed. Given the general nature of the book, much of this is discussed on a very broad level, to provide the reader with a guide of where to look and what to look for when thinking about the legal aspects of data storage. Whether it's data protection laws, encryption technology laws, or jurisdictional issues, (or more!) this section may well be especially relevant to the non-lawyers among us who are looking for a broad overview of what they should have in mind.

Useful resources

The Frontline SMS Users' Guide to Data Integrity
http://www.frontlinesms.com/wp-content/uploads/2011/08/frontlinesms_userguide.pdf

Deflect *https://deflect.ca/*

Cloudflare *https://www.cloudflare.com*

For a listing of secure tools, see *https://www.prismbreak.org*

On choosing a hosting provider,
see *https://learn.equalit.ie/wiki/Responsible_Data_Forum_on_Hosting*

On setting up a secure hosting provider,
see *https://learn.equalit.ie/wiki/Secure_hosting_guide*

NGO Law Monitor - *http://www.icnl.org/research/monitor/*

Maps of Data Protection Laws
http://www.forrestertools.com/heatmap/
& *http://www.dlapiperdataprotection.com/#handbook/world-map-section*

Choosing an Open Source Licence for Code
http://choosealicense.comCreative Commons - https://creativecommons.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.

RESPONSIBLE DATAFORUM

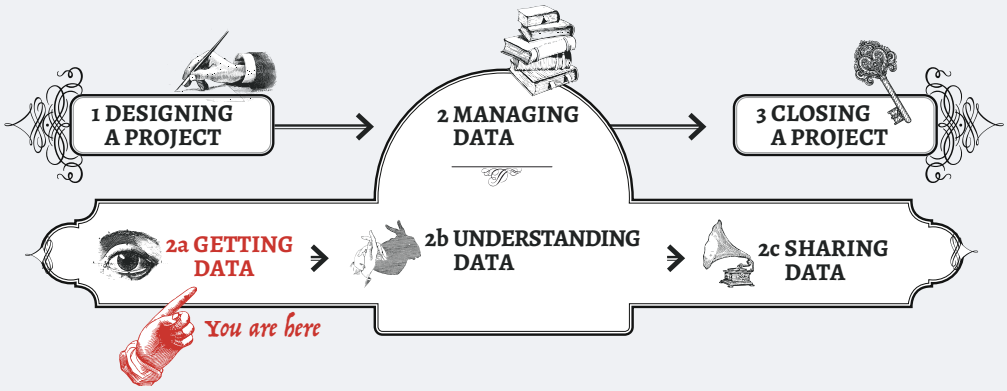
The HAND-BOOK
OF THE MODERN DEVELOPMENT SPECIALIST

2a

GETTING DATA

COLLECTING DATA FOR YOUR PROJECT

The project data flow



Target audience

Project planners who are getting down to the data details; thinking through what kinds of data the project will need and how to get that data. Issues associated with using datasets from other people are addressed here, as well as things to consider when collecting new data, respecting *data minimisation* principles. Finally, a comprehensive section on *consent* might be of use for those who are wondering what *meaningful consent* actually looks like.

When might this chapter be useful?

When you're getting down into the details of collecting data for your project.

Content summary

Why are you collecting data in the first place? To make sure your data gathering is relevant and actually useful, this section suggests framing your data collection in terms of a concrete question and building out processes and data points based on that specific question. You might not always need to directly collect your own data, either, as there are ways of getting data from existing sources - such as Freedom of Information requests from governments, or simply searching to see if anyone has already published the data you need.

Collecting new data

In the case that you do need to collect your own data, a list of questions is provided to help you plan a responsible data collection process. An exercise is suggested to help you 'define challenge areas' with your team in advance - for this, having a diverse group of people and perspectives will help strengthen the answers.

With responsible data practices in mind, we encourage collecting the minimum data required to answer your question. Practising *data minimisation* can help ensure that resources aren't wasted and that unnecessary (or potentially harmful) data isn't accidentally collected. Broadly speaking, there are a few 'red lines' that you should think carefully about when deciding to collect a new dataset.

LIVE DATA PROCESSES

Systems which collect and process data on an ongoing basis can be the most tricky to plan, as they compress the data cycle, and mean that individual team members might need to play multiple roles. A '*Roles-Responsibilities-Functionality Map*' tool is suggested here to help plan ahead and clearly assign roles and responsibilities within a certain team.

Next, various ways of collecting data are considered, followed by a checklist for tools, security and training that might help think through various technology-related considerations when setting up a data collection process.

WORKING WITH EXISTING DATA

If you can identify existing datasets that support your project activity, there are three primary areas of responsibility to consider. They are explored in more detail here:

1. *How the data is evaluated for use*

Understanding the biases that have gone into the dataset is crucial to making sure it is used in an appropriate and responsible way. Sometimes this is easier said than done, so to help think through these considerations, and gain appropriate contextual information around the data, a short list of questions to go through with your team is included here. You might not need to find the answer to all of them, but actively deciding which of these is most relevant for your project will help in this process.

2. *How it is managed*

Best practices with regards to data management are explored here, such as keeping multiple, well-labelled versions of your data if it is undergoing transformations. Some legal and licensing considerations are also outlined to make sure you respect the intellectual property restrictions that may (or may not) be associated with the data.

When manipulating the data, it's important to keep in mind *the people in the data* How might various manipulations or combinations of datasets reveal personal or private details about their lives?

Thinking through carefully what kinds of information could be revealed from a certain dataset is crucial, and can provide a useful, though perhaps dystopian, project planning exercise.

3. *How it is presented*

Including context to the data is vital, and so is being transparent about what the data does, and doesn't, represent. Being explicit about what is and isn't there can help others interpret and manage your data responsibly.

Power to the people

Ultimately, much of the data we are collecting refers to a particular group of people, and it's the people who are reflected in the data who should have the most say over how the data is used. In practice, making sure that this is the case can be trickier than it sounds, though, so here we include a number of considerations to think about to try and ensure that more power lies with those represented in the data, than those doing the collection. With regards to *representation*, for example, there are two main considerations: *inclusion* and *accuracy*.

Consent

Different types of consent are discussed in detail in this section. For those who have heard terms like "*informed consent*" used but not really understood what this means in practice or in legal terms, this section may well prove especially useful. Different required elements in a responsible consent procedure are considered here in detail, such as the items needed to be included in a consent agreement, issues that should be brought up with participants in the agreement, and ways of disclosing information to that participant or group of participants.

Developing a strong consent policy within a project or an organisation can help others set up responsible consent procedures more easily, so some components within a responsible consent policy are included here, along with advice on how to implement and maintain such a policy.

Finally, a word of caution is offered: is consent broken? A brief overview of this argument, and some ways to help create meaningful consent, are included here.

Useful resources

Data QualYtI: Do you trust your data?
(an article Hjusein Tjurkmen, Mariyana Hristova, Musala Soft)
http://istabg.org/data-qualyti-do-you-trust-your-data
more at *http://schoolofdata.org/handbook/courses/finding-data/#sthash.8o6YozUJ.dpuf*

Organizations that may be able to help you further on issues you may encounter:

Geeks without Bounds *gwob.org*

Datakind *www.datakind.org*

A checklist for evaluating policies on consent is available at
https://docs.google.com/a/theengineroom.org/document/d/1PJxBAP1rFkjq9p7NuYcN_G5iomfCML-qiMTn5SPPHxE/edit

Solove, Daniel J., Privacy Self-Management and the Consent Dilemma (November 4, 2012). 126 Harvard Law Review 1880 (2013); GWU Legal Studies Research Paper No. 2012-141; GWU Law School Public Law Research Paper No. 2012-141. Available at SSRN: *http://ssrn.com/abstract=2171018*

Launching an SMS code of conduct for Crisis Mapping
http://irevolution.net/2013/02/25/launching-sms-code-of-conduct/

the engine room

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

RESPONSIBLE DATAFORUM

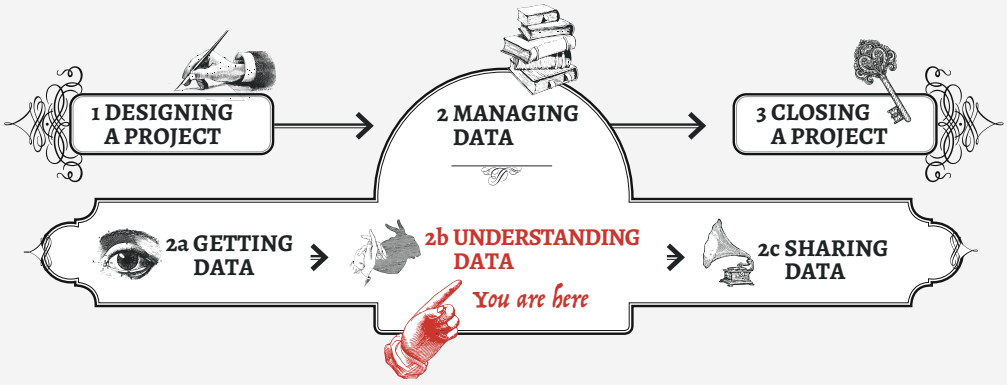
The HAND-BOOK
OF THE MODERN DEVELOPMENT SPECIALIST

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.

RESPONSIBLE DATAFORUM

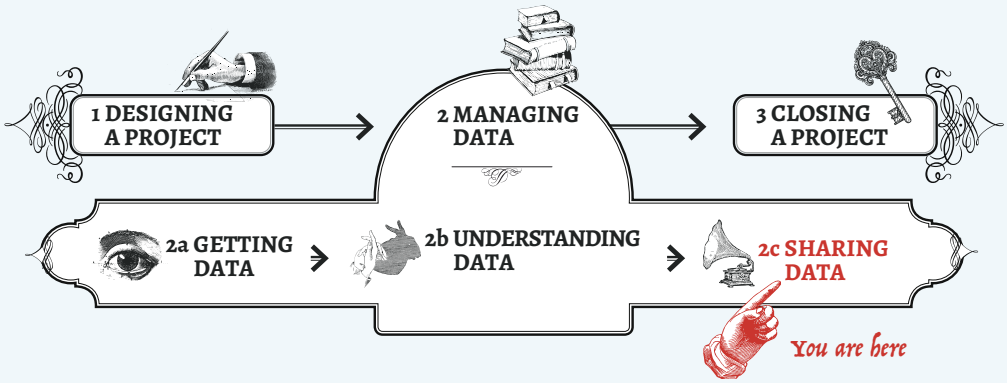
The HAND-BOOK
OF THE MODERN DEVELOPMENT SPECIALIST

2c

SHARING DATA

WHAT TO DO WITH YOUR PROCESSED DATA

The project data flow



Target audience

Those who are managing how the data will be dealt with once it has been analysed, making decisions on how it will be shared or published, the licenses that will be used, and how it might eventually get presented.

When might this chapter be useful?

When you're considering you have collected, or the data that you have analysed; is it appropriate to be made public, or shared with a smaller group of allies?

Content summary

The benefits and limitations of sharing or publishing data are discussed in detail here. Being 'open by default' might be useful in some situations, but when working with sensitive data, it's better to err on the side of prudence before making absolutely sure that no sensitive or potentially harmful data is being released into the open.

The issue of *data ownership* is a tricky one; to whom does the data belong? Those who are represented in the data, or those who have collected the data? There's no easy answer to this question, so thinking it through carefully on a project-by-project basis is necessary to make sure you come to an appropriate answer.

If you do decide to share it, making sure it has an appropriate license which represents well what you want the data to be used for can help clear up any uncertainties or potential misuses of the

data. If you're sharing it with a specific group of actors, a legal or contractual framework might help to clearly outline what the data can and can't be used for by other parties, and different frameworks for establishing these principles are discussed here.

Sharing can happen on different levels; within an organisation, within a closed but controlled set of actors, or - the point of no return - freely shared online. Different pros, cons and considerations for each of these three groups are discussed here, and might be helpful if you're not totally sure which level would be most appropriate for you.

PUBLISHING DATA

If you decide to share it online, there are a number of further things to consider; do you want it to be *open* or *closed*? Chances are that if the data is suitable to be online, it's suitable to be used by others - but this isn't always the case.

If you want to make your data *open* then some tips are included here; like how to give it an *open* license, or make it fit within the *open definition*, or publishing it to the internationally-recognised transparency standard, the International Aid Transparency Initiative.

Before publishing though, you might want to make sure that individuals are taken out of the data - but this isn't always so easy. A section discussing *anonymisation* - and various strategies for *de-identification* of data are included next.

Anonymisation and identification is a particularly difficult responsible data issue, as 'perfect' anonymisation is difficult if not impossible. Releasing personally-identifiable information can be harmful to the people to whom the data pertains, but it can also be hard to remove and make sure that at no point in the future it will be re-attributable.

Some suggested strategies and techniques are offered here, but with the caveat that if you're particularly worried about identification of your data before publishing, it might be most useful to simply get help from those with more experience in anonymisation and de-identification.

PRESENTING DATA

No data is neutral, and normally when you present data you want to convey a specific narrative. Doing this in a way that doesn't skew the data or misrepresent the topics in question is important not just for the reputation of your organisation, but also as an ethical responsibility to the people who are reflected in the data itself.

This section offers a very brief overview of "*do's*" and "*don'ts*" when it comes to presenting your data, though diving into the details around ethically and responsibly visualising data lies way beyond the scope of this book. For more on this topic, we suggest referring to the accompanying suggested resources.

Useful resources

'License Chooser' which will help you pick the right license for you: *https://creativecommons.org/choose/*

Frequently Asked Questions about the Open Data Commons licenses: *http://opendatacommons.org/faq/licenses/#General*

More information on Open Data Licensing: *http://opendefinition.org/guide/data/*

More information about IATI: *http://www.aidtransparency.net/about*

The UK Anonymisation Network provides consultation and training for NGOs. Anonymization guide produced by the UK Information Commissioner's Office *http://ico.org.uk/~/media/documents/library/Data_Protection/Practical_application/anonymisation-codev2.pdf*

Mushon Zer Aviv's short essay "How to lie with data visualisation": *https://visualisingadvocacy.org/blog/disinformation-visualization-how-lie-datavis*

TacticalTech's Visualising Information for Advocacy guide: *http://visualisingadvocacy.org*

School of Data's Visualizing Data course: *http://schoolofdata.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.

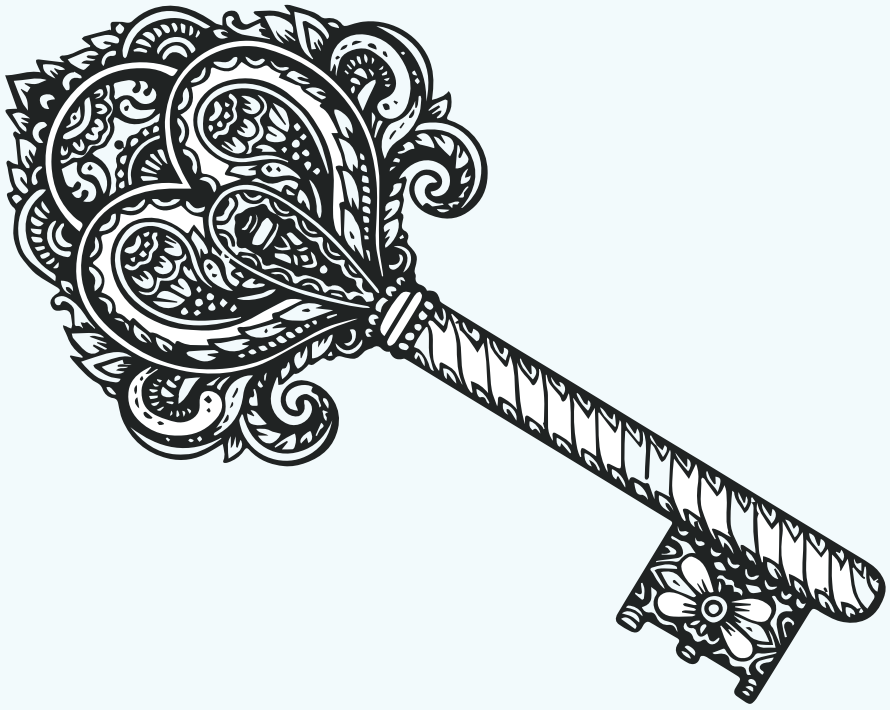
RESPONSIBLE DATAFORUM

The HAND-BOOK
OF THE MODERN DEVELOPMENT SPECIALIST

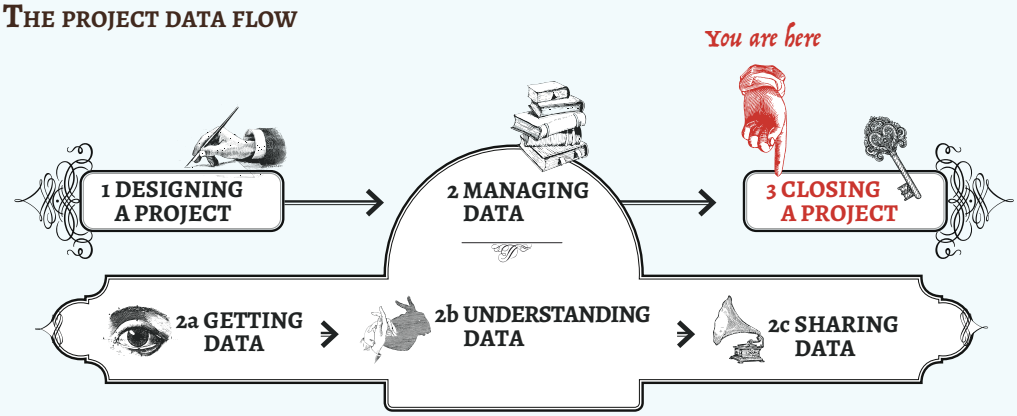
3

CLOSING A PROJECT

ARCHIVING AND PRESERVATION OF CONTENT



The project data flow



Target audience

Project managers who are coming towards the end of a project.

When might this chapter be useful?

When you're approaching the end of the project, but still have a little time left on it for some final decisions around archiving and preservation of content.

Content summary

This chapter addresses an oft-forgotten part of the project: the end of it. Too often, the data that is produced during a project is left to one side during the rush of ending a project, but in this section we suggest that deciding what to do with the data after it has served its purpose is an integral part of the project design process.

To plan for project closure, start by thinking through the types of data you have, and where it all lives. Then you can decide what to do with it; whether that is *disposing of it* (and making sure that that is really the right decision for you, and that the data isn't being used by others) - or working out a responsible way of *archiving* your data. Preserving your information has the potential to be used by others, as well as bringing historical value.

If the data in question is very large, substantial data infrastructure may well need to be in place to support it being archived. If this is too much of a technical burden, then some potential partners for archiving your data are suggested here to make this process a little easier.

Though it might be easy to forget, what happens to the data at the end of a project can make a huge difference to how the project is referenced, how its legacy lives on (to inform other projects, or not) - and, most crucially, to how the people referenced in the data are treated.

Useful resources

Archivists' Guide to Archiving Video (WITNESS) *http://archiveguide.witness.org/* and a video on this topic *http://blog.witness.org/2014/10/video-series-archiving-and-preservation-activists/*

New Tactics online discussion on archiving for human rights advocacy, justice and memory *https://www.newtactics.org/conversation/archiving-human-rights-advocacy-justice-and-memory*

The Open Society Archives *http://www.osaarchivum.org/*

Duke University Human Rights Archive *http://library.duke.edu/rubenstein/human-rights/*

Human Rights Web Archive at Columbia University *http://hrwa.cul.columbia.edu/*

University of Texas Libraries' Human Rights Documentation Initiative *http://www.lib.utexas.edu/hrdi*

Archivists without Borders *http://www.arxivers.org/en/asf_internacional.php*

the engine room

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