How to do data-driven investigations in 12 memes

(Even though you're not an investigative journalist)

You're not a journalist (at least not in the traditional sense), but you care about certain issues and the internet hosts an abundance of information. Is it possible to do a data-driven investigation anyway?

We examined this question during our workshop (titled "Who reads open secrets?") at UN|COMMONS, Berliner Gazette's 15th annual conference. Based on some of our experience on working with data, we came up with 12 simple steps (and memes!) which can guide non-data journalists through some of the basics on how to do data-driven investigations.

These steps/memes are designed for anyone who's wondering how to start-off with a data-driven investigation and only include some of the very basics. More comprehensive information is included in various other resources, such as the Data Journalism Handbook. Additional resources on searching for information, scraping websites, processing, analyzing and visualizing data are included in the "resources" section at the end of this blog. But before you get started, it's important to evaluate risks and to take precautions.

Threat model

Doing an investigation can potentially be risky, depending on your threat model (the types of risks that you might face based on your environment). Various governments around the world, for example, are notorious for cracking down on journalists and activists. You might not consider yourself as one, but the parties that you are investigating might view you as a threat to their interests nonetheless. As those in power often have low tolerance for investigations into their affairs, doing a data-driven investigation might be risker than what you might imagine.

So how to assess your threat model? You can start off by checking out the Electronic Frontier Foundation's (EFF) Surveillance Self-Defense guide which includes the following questions worth considering:

- What do you want to protect?
- Who do you want to protect it from?
- How likely is it that you will need to protect it?
- How bad are the consequences if you fail?
- How much trouble are you willing to go through in order to try to prevent those?

While the EFF's guide explains how to answer the above questions, it might also be worth seeking *legal advice* which is specific to your investigation.

Digital Security

Protecting your data and that of your sources through the use of *encryption* is good practice because it provides confidentiality. By encrypting your emails, phone calls, instant messages, files and hard drives, not only do you protect your sources, but you also reduce the probability of third parties (perhaps including your investigation target) figuring out what you're up to.

Fortunately, basic digital security tools which encrypt files and communications don't require “rocket science” and anyone can learn to use them. Security in-a-Box for example - Tactical Tech's hands-on guide - provides detailed steps on how to use various tools and tactics for digital security. Additionally, CryptoParties are constantly held in various cities around the world, where everyone is welcome to attend and to learn in a hands-on way how to use basic cryptography tools. However, you might want to check whether the use of encryption is legally prohibited in your country and, if so, evaluate the risks.

Online anonymity might also be essential to your investigation, depending on your threat model – as explained above. Tor software provides online anonymity and can help you gain access to blocked websites.

Resources

As promised, we've included more resources at the end of our blog which can help you get started with your data-driven investigation.

Accessing Data:

http://wikileaks.org/
http://www.cryptome.org/
https://publicintelligence.net/
http://governmentattic.com/
https://www.muckrock.com/
https://www.foiamachine.org/
http://www.data.gov/
https://archive.org/index.php
https://www.wikidata.org/wiki/Wikidata:Main_Page
https://www.wikipedia.org/
http://www.pacer.gov/
https://www.wolframalpha.com/
https://www.aclu.org/
http://www.theguardian.com/world/the-nsa-files

https://leaksource.wordpress.com/category/nsa-files/
http://datacatalogs.org/
http://data.gov.uk/
http://www.who.int/research/en/
http://oad.simmons.edu/oadwiki/Data_repositories
http://datahub.io/
www.openstreetmap.org/
https://code.google.com/p/worlddb/
http://geocommons.com/
http://opencorporates.com/
http://www.icharts.net/

Scraping/APIs/Processing Data:

https://import.io/
https://scraperwiki.com/
http://sunlightlabs.github.io/datacommons/
http://sunlightlabs.github.io/openstates-api/
http://sunlightlabs.github.io/Capitol-Words/
http://sunlightlabs.github.io/partytime-docs/
https://dev.twitter.com/
https://github.com/rhodey/LiberDat
http://data.worldbank.org/developers/
http://opendata.socrata.com/api/docs
http://project-open-data.github.io/
http://open311.org/
https://www.census.gov/developers/
https://www.cometdocs.com/

Processing/Filtering Data:

https://www.elasticsearch.com/
https://documentcloud.github.io/docsplit/
http://nokogiri.org/
https://github.com/onyxfish/csvkit
http://tabula.nerdpower.org/

Data Extraction:

https://github.com/louismullie/treat
http://www.textteaser.com/
http://nlp.stanford.edu/software/corenlp.shtml
http://www.alchemyapi.com/
http://www.transana.org/
https://www.opencalais.com/

http://www.datasciencetoolkit.org/

Data Aggregation/Analysis:

https://www.documentcloud.org/home
https://github.com/documentcloud/docsplit
http://www.civomega.com/
http://www.r-project.org/
http://josephwilk.github.io/rsemantic/
http://mallet.cs.umass.edu/topics.php
http://nlp.stanford.edu/software/tmt/tmt-0.4/
https://www.overviewproject.org/
http://granoproject.org/
http://tables.googlelabs.com/

Data Visualization/Storytelling:

http://timeline.knightlab.com/
https://datawrapper.de/
https://www.paterva.com/
https://github.com/glejeune/Ruby-Graphviz
https://gephi.org/
http://igraph.sourceforge.net/
https://oicweave.org/
https://developers.google.com/chart/
http://leafletjs.com/
https://maps.google.com/
https://github.com/wlwardiary/cable2graph
https://infogr.am/
http://d3js.org/
http://www.flotcharts.org/
https://www.tableausoftware.com/public
http://www-958.ibm.com/software/data/cognos/manyeyes/
http://raphaeljs.com/
https://www.mapbox.com/
https://piktochart.com/

Workshop team: Marie Gutbub (Centre for Investigative Journalism), Maria Xynou (Tactical Technology Collective), M.C. McGrath (Transparency Toolkit), Valentina Pavel (ApTI), Christoph Zeiher (Zeit Online)