Chapter Four - The Data Infrastructure

Simona Simona

4/12/2020

In this chapter we cover the following topics:

- What is social science data
- Where do I find data for my research
- Primary vs. secondary data
- Third party data collection
- Webscrapping
- Social media data mining
- Database integration with R
- Secondary data infrastructure
- Online experiments

What is social science data?

The concept of data in quantitave social science was very straightforward not long ago. Researchers came up with a topic based on their interest to fill the gap in knowledge that exist in their area of interest and then came up with data collection tools, involving a survey in most cases and went on to collect data themselves. However, we are now living in what is called the 'age of data' and it has revolutionalised what is possible in all relms of life including social science research. The move from analogy to the digital age has come with it a shift in people's way of life. Many people's daily lives which used to be run in analogy are now digital. Things like buying, shopping, selling, banking, entertainment, music, education etc. This transition has brought with it an explosion of data that has come to be known as 'big data'. Literally everything we do online is captured and stored somewhere whether we know it or not. This trend is increasing an an astonishing speed.

This means social science researchers have an opportunity to access an enormous amount of data to do research but also to use the parallel growth in computing power to conduct online studies that were not possible to do before. It is worthy noting however, that the data explosition is in many respects not be structured in a tradional format that social scientists are used with. It requires strong data management, analytic and computing skills that social scientists that are not part of formal training that social scientists get in undergraduate and graduate schools. New disciplines have emerged specifically to liverage these opportunities including data science, machine learning, artificial intelligence, big data analytics, deep learning, computational social science etc. Social scientists are highly represented in these fields although most backgrond training of scholars involve a blend of computer science, information technology and statistics.

Social scintists are not completely sold by the digital data hype because of lack of skills and the problems that come with the data itself. These include lack of representiveness, messiness, incompleteness, inaccessible and drifting[1]. Which means that although some quantitative social scientists are involved in these new innovative data initiatives, the large majority are still involved with traditional research. What seem to have changed though is an increased use of secondary data compared to primary data.

In this chapter I try to represent the social science data infrastructure to suit the current trends in quantitative social science. I will briefly discuss sourcing primary data in all its manifestations, then move on to the secondary data infrastruture including big data platforms commonly used in social scinces. It is not possible to provides recipes for collecting data in all cases. I try to give an overview of data sources and in some cases prescribe computational methods of data collecting for those that require such methods.

Primary data sources

Primary data is the data collected by researchers themselves using different methods. They may use surveys, interviews, observations or experiments. **Surveys** are quantitative methods involving a list of questions aimed at extracting certain kinds of information from a specific group of people. Data is measured by respondents's self-reports about their own behaviour, feelings, actions and thoughts. They can be conducted via face-to-face, phone, email or over the internet. **Interviews** involve one-on-one or small group question and answer sessions. The researcher often comes up with a series of questions to be followed strictly or used as guides for the interview sessions. They are used when exaxt knowledgeable opinion on the research question is desired. **Observations** can be long-term or shot-term but they involved detailed monitoring of a group of people, event or locality. They are used when researchers want to avoid biases that may intrude in an interview or survey. **Experiments** could be either true of quasi. They involve random allocation of participants into experimental and control group and administering an intervention with the two groups being compared in the end to see if the intervention is responsible for any change in the outcome. Experiments are usually preferred when the researcher wishes to determine cause and effect.

There are certain reasons that may particularly compel social scientists to conduct primary studies. Firstly when the problem we are working on is noval with no background information information. As I am writing this part, I am in self-quarantine during the coronavirus pandemic in 2020. Supposed, I wanted to research the economic implications of the coronavirus (covid-19) pandemic on an indigenous population in North Africa. Even if I may have access to previous studies that have looked at the impact of other pandemic situations or even other coronaviruses, I obviously do not have any reference materials on covid-19 because this is the first time it is occuring. Find my research findings are to be credible, I would need to conduct a primary study, produce new data and analyse it.

Secondly, when we are working on a specific group of people. If my study looks at the use of traditional medicine among the lozi people of Barotseland in western Zambia, it would be difficult for me to find already existing data that addresses this problem. I would ndeed to design a new study that requires the collection of new data from western province in Zambia.

Thirdly, when we want to confirm or dispute national results with local trends. If for example you think that the national crime prevalent rates do not represent a certain locality, you would implement a local study that looks at crime rates and compare them to the national statistics. This study may replicate a national study only this time it is conducted among the local population

[1] Big data systems are changing all the time in terms of the systems themselves, who is using them and how people are using them. This offers a unique difficulty in terms of studying long-term trends ()

Fourthly, solving a local problem. If we want solve a problem which is specific to a certain area, we would conduct a study in the concerned locality. Previous research may not help us here especially if the problem is noval. For example, if you have serious drought conditions in certain parts of the country and you are conducting a study that seeks to address the impact of the drougnt condition on food security. We would ned to conduct a primary study to address the problem.

Challenges of primary data

There are a few factors which may make quantitative social science researchers opt out on using primary data. Firstly, resource constraints. Primary data collection requires several pieces to come together and these may not necessarily be cost effective. You need well trained research assistants, ethical approval, data quality issues, The problem with primary data is that it is neither cost-effective nor time-effecient

Third party primary data collection

The problem with secondary data is that it was not mearnt for research therefore it takes a huge amount of time to repurpose it to your desires. It should be one of the reasons you are studying this book because we are going to move from acquiring data through data manipulation using different methods up to data analysis and visualisation

Web applications

Data from Twitter

In this post, I document the process and tools used to acquire data from the Twitter API. I will use the R statistical programming language through RStudio in this process. It is important that you install R and RStudio on your machine, including all the accompanying packages specified here. Get documentation providing installation guides for both R and RStudio here. We are going to use the RStudio environment for our data collection but of course it will not work properly if you haven't installed R first as he installation document explains. Assuming everything is set, we will need the following suite of packages. We might as well begin by installing them.

```r
install.packages("twitteR") # As the name indicate, it is used for capturing data from twitter
install.packages("ROAuth") # You need to be authorised to get data from twitter, this package
install.packages("httr") # For requesting data from APIs
install.packages("rtweet") # Another package we use to get data from twitter and analyse it
library(twitteR)
library(ROAuth)
library(httr)
library(rtweet)
```

After we are done with the packages, we will need credentials authorising us to access twitter data. It is important for us to have credentials so that twitter can control how much information we are gathering and that we do not abuse the platform. To do that, we need to create an app on the twitter developer site and we will use this app to call the twitter API. Click here to access the website or you can just Google twitter developer platform or any combination of words which have the words twitter and developer next to each other. We need to have a twitter for us to do this, which I assume we already have, otherwise why would we need to extract data from a community we don't belong to. On the developer platform, we will create a developer account (yes it is different from your main twitter account). Twitter will ask for your mobile number here. It is important for you to supply the correct number because they will use it for authentication. Click on the 'Create New App' once you sign in. It is located on the top right corner of the screen. Name and describe the app as you wish. The instructions are pretty straightforward. Some fields are mandatory. You will be requested to provide a website and callback URLs. You can provide your own website if you have one. Any website should suffice here. The callback URLs are not mandatory – you can leave them blank.

After the app is created, it will appear on your screen and you should be able to click on the 'details' button, to get a page that looks like a picture below. You can see the name of my app, description, my website URL and callback URL as indicated above. On the 'keys and tokens' tab, you will find the credentials that are needed for you to be authorised to collect data on twitter. I have chosen not to show that tab here because it contains my own credentials and I didn't want them exposed to anybody who reads this post. You also need to keep yours confidential because anyone can (mis)use them if they are exposed. They include API key, API secret key, Access token and Access token secret.

Now that we have our credentials in the bag, there are basically two packages in R (at least those I know of this far) we can use to harvest data from twitter: the rtweet package by Michael Kearney and the twitteR package by Jeff Gentry. Either should do the job perfectly well. I have included the procedure for both just in case one of them gives you problems. It is certainly not uncommon for packages to give you a few glitches here and there even if you are an experienced user.

For the rtweet package, we need to copy the credentials and put then in the create_token function in R. Please note that the credentials I am using here are only for illustrative purposes. But your credentials should look more or less like them.

```
token <- create_token(
  app = "simona_app",
  consumer_key = "VocbltyHiK5F2VHnC4jRpmaT8",
  consumer_secret = "wrwVFnoaRXZUYsCfr0JvEgUcZmKLsHasPQjVgz6HJHlTaDeM7uRl",
  access_token = "786205049854307650-BvJaX0YUfVtDz0wfAslGyxRhr43Y9M",
  access_secret = "y5Dxe9YsCOLKdCTOyRW6btR5KyTAHlssEN3WFSBvGjYtu")
```

For the twitteR package, you just need the keys to be in the following format. The exact wording of the access token shouldn't matter much.

```
api_key <- "VocbltyHiK5F2VHnC4jRpmaT8"
api_secret <- "wrwVFnoaRXZUYsCfr0JvEgUcZmKLsHasPQjVgz6HJHlTaDeM7uRl"
access_token <- "786205049854307650-BvJaX0YUfVtDz0wfAslGyxRhr43Y9M"
access_token_secret <- "y5Dxe9YsCOLKdCTOyRW6btR5KyTAHlssEN3WFSBvGjYtu"
setup_twitter_oauth(api_key, api_secret, access_token, access_token_secret)
```

Assuming I want to examine gender-based violence discourses in virtual spaces and I want to collect 5000 tweets which make reference to gender-based violence. The number is arbitrary, you can choose your own sample size here if you want depending on what you want to study. Below we specify the commands for requesting data from the twitter API using our credentials. For the twitteR package the key objects should appear in your global environment. The retweet package will only show the token object and environment in the variables section.

```
tweets <- search_tweets("gender violence", n = 5000, include_rts = FALSE )
```

The code from rtweet is pretty straightforward. It says we are looking for 5000 tweets which contain gender and violence anywhere in the tweets and we want our search results to be put in the object we have named tweets. The search_tweets() is the function which performs the job. The include_rts=FALSE bit means we don't want retweets to be included in our search. Specifying the same query in the twitteR package, our code would look like the following. Not so different: We want 5000 tweets containing the words gender and violence of any language and we want them put in the tweets object. Note that here we have not told the twitter API to exclude retweets.

```
tweets <- searchTwitter("gender+violence", n=5000, lang=NULL)
```

We should now have our data and the next job, which should be the subject of my next post, is to manipulate it so that it can be ready to be used in quantitative text analysis. But before we go, I want to introduce another platform, which you can use to mine twitter data, especially for those who get a little bogged down by programming stuff. It has search and visualisation capabilities built within it. It is called COSMOS, developed by the social data science lab at Cardiff University. It is free for academic institutions and not-for-profit organisations. The interface looks as follows and you will find demos on the website on how to collect and analyse the data. It is pretty straight forward and I highly recommend it.

```
## Put the COSMOS here!
```

Airbnb

Amazon

FACEBOOK

YouTube

Secondary data infrastructure

The Demographic and Health Survey (DHS)

The Demographic and Health Surveys (DHS) are nationally representative population-based cross sectional survey of women and men of reproductive age (15-49 year for women and 15-59 years for men) designed to provide information on a number of measures including fertility, family planning, mortality, nutrition, maternal and child health, HIV/AIDS, domestic violence and other health indicators, at national level for both rural and urban areas of DHS countries. The DHS are funded mainly by the United States Aid for International Development (USAID) in collaboration with other donors and participating countries. Technical support is provided by the DHS program, formally MEASURE DHS. The DHS program is implemented by ICF International and have since 2013 been joined by partners Blue Raster, The Futures Institute, Johns Hopkins Bloomberg School of Public Health Centre for Communication Programs (JHUCCP), PATH, and Vysnova, EnCompass and Kimetrica. The DHS program has provided technical assistance to more than 300 surveys in more than 90 countries since the DHS inception in 1984 (DHS Program, 2018).

The main purpose of the DHS is to provide policy-makers, program planners and researchers in DHS participating countries with a database sufficient to allow them to make informed policy and program choices; to expand the international health and population databases; to advance survey research methodology for the collection and processing of demographic and health data; and to help participating countries develop the technical skills and resources necessary for conducting their own demographic and health surveys (Fabic et al., 2012; Fisher and Way, 1988).

There are different types of surveys implemented by the DHS program such as the AIDs Indicator Survey (AIS), Malaria Indicator Surveys (MIS), Service Provision Surveys (SPA) and the Key Indicator Surveys (KIS) (DHS program, 2018). this study only makes use of the standard DHS (to be interchangeably referred to simply as DHS in this study). The DHS are the largest and most comprehensive health surveys in sub-Saharan Africa. Their most important characteristics relevant for this study is adherence to standardised sampling design, questionnaire construction and implementation procedures across all participating countries. This characteristic enables the surveys to be comparable across countries and over time.

The DHS data for each country are stored on the DHS program website . The data for each country is stored using six files which contain Individual, man, birth, couple, child and household recodes.The DHS sample is typically representative at national level, for urban and rural areas, the regional level and sometime at state/provincial or district level. The surveys have large sample sizes (usually between 5,000 and 30,000 households) and are typically conducted about every 5 years (DHS Program, 2018). The standard DHS collects data using four core questionnaires: household, individual women, individual men and biomarker questionnaires. A household questionnaire is used to collect information about characteristics of the household's dwelling units and demographic profile of every usual member of the household and visitors. Information such as age, sex, relationship to the head of the household, education, parental survivorship, residence and birth registration is collected (DHS Program, 2018). It is also from this questionnaire that members of the household who are eligible for individual interviews using an Individual Woman's or Man's Questionnaire are identified.

World Development Indicators (WDI)

World Development Indicators (WDI) are the primary World Bank collection of development indicators, compiled from officially recognised international sources. These are the most current and accurate global development data available and they provide national, regional and global estimates (The World Bank, 2018). The WDI database is released annually by the World Bank late each year, it contains statistics up through the previous calendar year and is freely available online along with supporting interpretive book of key tables (Babones, 2014). As of 2018, the statistical reference includes data on around 1,400 indicators covering more than 220 countries and territories with some data indicators extending up to Table 5.1: Description of Sample Data of DHS 2006-2015 in sub-Saharan Africa 50 years back. The main indicators covered include economic policy, debt, finance, education, infrastructure, labour and social protection, environment, poverty, health, public and private and trade (Babones, 2013).

The WDI primarily organises countries according to income levels and these are low, lower middle, middle, upper middle and high. The database also designates all countries of the world one of the six regional labels of the World Bank (in the parenthesis are numbers of countries categorised as such): North America (3), Latin America and Caribbean (42), East Asia and Pacific (37), Middle East and North Africa (21), Europe and Central Asia (58) and Sub-Saharan Africa (48).If your study is focussed on acertain regin of the world, you could concentrate on that.

Freedom house

Freedom House is a non-governmental organisation that has been publishing a freedom in the world report on the state political rights and civil liberties for over 190 countries since the early 1970s . Their methodology mirrors the 1948 Universal Declaration of Human Rights (UDHR), premised on the universality of standards of civil and political across the world – that these standards apply to all countries and territories irrespective of geographical location, ethnic or religious composition or level of development (Freedom House, 2019).

The freedom score is based on the events and activities happening in a particular country for the concerned time period. The score is arrived at by consensus and deliberated over a series of meetings involving of more than 130 analysts, advisers and staff with a global representation. They use a suite of data sources including newspapers, academic research, NGO reports, professional contact and on-the-ground research. An element of subjectivity may be unavoidable, but the ratings process emphasises methodological consistency, intellectual rigour, balanced and unbiased judgement. The fact that it has been used in many studies also shows their trustworthiness among academic scholars.

Freedom status ratings are derived from 25 questions representing political rights and civil liberties. The questions address electoral process, political pluralism and participation, functioning of government, freedom of expression and belief, rule of law, associational and organisational rights and personal autonomy and individual rights (Freedom House, 2019). The overall scores of both political rights and civil liberties add up 100 points . This study uses these overall average ratings of 1 to 100 to represent the degree of freedom or freedom status in a country. In the empirical analysis, they are standardised with the mean of 0 and standard deviation of 1 for easy comparability and interpretations as indicated below. The freedom status variable in this study represents civil and political liberties and is applied in chapter seven which looks at the influence of civil liberties and socioeconomic entitlements on maternal health care.

World Governance Indicators (WGI)

The World Government Indicators (WGI) are research datasets summarising cross-country indicators of the quality of governance from 31 different sources capturing governance perceptions from non-governmental organisations, commercial business information providers, public sector organisations, surveys of households and firms worldwide . The WGI are implemented by the World Bank and consist of six composite indicators of governance including voice and accountability, governance effectiveness, political stability, regulatory quality, rule of law and control of corruption (Kaufmann et al., 2011). They use unobserved components model statistical methodology to standardise the data from different sources to make it comparable and then aggregate weighted averages of the individual source variables to create composite indicators. Margins of error are also constructed to reflect the imprecision inevitable in governance measurements (Kaufmann et al., 2011).

KOF Globalisation Index

The KOF Globalisation Index is housed by the KOF Swiss Economic Institute. It was first introduced by Dreher (2006) and updated in Dreher et al. (2008). It measures globalization along the economic, social and political dimension for almost every country in the world since 1970. It now gets released on a yearly basis. The details of what the dimensions of globalisation are measuring are included in chapter eight where is data is used. The methods for calculating the indices are complex but they are available on the methods documentation on the Swiss Economic Institute Website.

I use this data to measure globalisation in this study because it is comprehensive, with more than 40 variables for almost all countries in the world. It is publicly available, and it is the most widely used globalization index in the academic literature (Potrafke 2015). A list documenting studies that use the KOF Globalisation Index is available at http://globalization.kof.ethz.ch/papers/. The data used in the study covered the period between 1970 and 2016.