CHAPTER II

# Sourcing Data

**The Research Basics of Sourcing Data**

1. Determine the question you are trying to answer.
2. Seek credible sources.
3. Know how to interview.
4. Understand the format of the data files.

*1. Determine the question you are trying to answer.*

*Who is your audience?*

*Why is the audience interested in this data?*

*What data is available?*

For example, university students, government officials, or journalists might want to know about dams in Brazil. A first question they may want to know is “How many dams are there?” Put this into context with the rest of the world - is it a relatively large or small number compared with other countries? How many dams are completed and how many are in progress? What are the effects of the dams? On local communities? On energy needs/expenditures? More will be discussed in latter sections on narrating your data, but you need to have a question in mind that you want to try to answer.

*2. Seek credible sources.*

*Primary v. Secondary* - Many times an article you read online will have been adapted from another article, or is a summary of data recently published online. Read through the article carefully and note any references to other articles, websites, or specific individuals that the article was getting its information from. Seek out the original source of information.

*Accurate* - This means no partial or biased data. Who is your source here? Sometimes organizations publish data that push its own agenda, or leave out information that may be potentially harmful to its reputation or the message it supports. Critiquing the data you find through academic sources is helpful because the author(s) often addresses all sides of biases. Be wary of data that looks too good to be true or that is published by institutions that have an explicit agenda.

*Triangulate -* This is standard in reporting. Triangulating your information means finding three credible sources that can independently attest to the information or data’s validity.

*3. Know how to interview.*

This is especially important if the data you are seeking to source is not already available. There are several types of in-person interviews:

*Structured -* Questions are recited exactly as they are prepared, and cues such as body

language, awareness of non-prompting, and other verbal and behavioral cues are noted.

*Semi-Structured -* Most popular. Interviewers follow question guidelines, however not

word by word. They may paraphrase or improvise based on the direction the meeting/interview is taking. General goals, questions, and desired topics are understood before the interview.

*Unstructured -* Most informal. These interviews are typically resourced as conversations,

in which the interviewer may ask questions with no specific guideline to follow. Questions are generally open-ended and the interview is flexible as it can take any direction the two parties choose to follow. Sometimes, this is the best way to discover information that the interviewer may not have known to even ask.

*Focus Groups -* These are sessions of 2 or more persons in which the interviewer would

like to hear responses from community members, organization, or specific groups as a whole. Special attention should be paid to those who claim to speak on behalf of the group, talk over others, or those that are especially silent. It is the responsibility of the moderator to diplomatically include everyone in the discussion, as well as take notes on what is said, not said, and other behavioral cues.

*4. Understand the format of the data files.*

Data is stored and presented in a variety of file formats. Some formats are easy for humans to read, but difficult for computers, while others are easy for computers to read, but may initially look incomprehensible to a human. Understanding the file format of your data will help you know what you can do with the data and if you need to take steps to save the data in a different format before using it. Some data sources are more “open” than others, allowing more people to access the data and use it for their own analysis. The following are common formats found while conducting research:

*PDF (.pdf) -* Portable document format files contain a fixed-layout flat document that includes the text, images, and graphics contained in the original. Saving as a .pdf allows the file to be opened on any computer, regardless of the application, software, or operating system used to create the file. PDFs are valued for their readability and aesthetic display. Data contained within PDF files is difficult for computers to read, even though it is easy for humans to see. A program to extract data from PDFs (see *Scraping Data* section) will be needed to convert tables within a PDF to a more manipulable format. It is not recommended to save open data, such as data tables, in PDF format, although it is a nice format for publishing reports or articles.

*Excel (.xls or .xlsx) -* Microsoft Excel is a spreadsheet program that allows for recording data in rows, columns, and cells. The program can be used for calculations, formulas, basic graphs and visualization, pivot tables, and macro programming. Several spreadsheets can be saved within a single Excel workbook. This format is often used for presenting open data since it is widely understood by people and allows for basic analysis within the program, and can be easily imported into databases and other data visualization tools. However, it is important to note that Excel is not an open-source software.

*Comma-separated values (.csv) -*  This file type is a text file that contains information in a table structured format where the data is separated by commas. It is often confused with an Excel file since .csv files often open automatically in Excel on most computers. CSV files are useful when importing/exporting large tables between format types or databases (for example, when exporting a table from a database to a spreadsheet program that cannot directly read data from the database the data can be exported as .csv by the database and imported as a .csv by the spreadsheet program). When opening a .csv file in Excel, be aware that formulas and formatting that you add to the data will not be saved unless the spreadsheet is saved in an .xls format. Only a single sheet in a workbook application such as Excel can be saved as a .csv at a time.

*JPEG (.jpg)-* This is the most common file format for digital images. Saving a photo as a JPEG allows the user to choose a level of compression, trading off image size for image quality. Images can be compressed to 1/10 the size of the original data without perceptible loss in quality, making it an ideal format for web or email which require smaller file sizes. Most file types are “lossless”, meaning that when the file is reopened all the original data will remain the same. However, JPEG is a “lossy” format, meaning that when an image is compressed, the original data is modified and will lose some of its detail.

*Text (.txt) -* .txt files are considered a universal format since they can be opened and read by any text editing or word processing software. These files contain text without formatting (such as bold or italics) and generally match that of a system terminal or simple text editor. Text is generally stored in ASCII format (American Standard Code for Information Interchange) which is based on the English alphabet, although there are modern character encoding schemes used to support other languages that based on ASCII.

*XML (.xml) -*XML (eXtensible Markup Language) is a language for describing data and documents in a format that both humans and machines can read. XML is often used for the interchange of data over the internet; XML consists of a series of tags (e.g. <country>), with attributes and text associated with each tag. HTML, which is used to describe the contents of a webpage, uses a pre-defined set of XML tags, but there are no restrictions on the tag names in XML files used to describe datasets. An XML reader is required to analyze XML-formatted information and transfer it to an application for display and reading.

*JSON (.json) -* JSON stands for JavaScript Object Notation. It is used to transmit data from a server to an application, and can be used as an alternative to XML (some find it easier to understand than XML). It is considered a ‘lightweight’ data interchange format that is human-readable and text based. There are some websites, such as json-csv.com, that can convert JSON files to CSV files for easier viewing (See *Scraping Data* section).



*Example of JSON versus XML data interchange format. JSON on the left, XML on the right[[1]](#footnote-1)*

1. Shandra Locken, “The What, Why and How of JSON for EDI Integration Specialists,” Aurora EDI Alliance, 2013. [↑](#footnote-ref-1)