Data Discovery Transcripts

Module 1 Transcript

Welcome to Data Discovery with York University Libraries. You are watching Module 1: "Introduction to data and statistics".

In this module, you will learn to understand when and why to use data and statistics, to differentiate between data and statistics, to classify different types of data, and to describe different methods of data collection.

So why do we use data and statistics in research? We can use data and statistics to answer a factual question. It can also be used to tell a story, in the form of a chart, graph, map, or data visualization, or to add clarity to a challenging or complex topic.

You can use data and statistics to explore relationships between two or more concepts or ideas, or for discovering patterns in information. You can use data algorithms to automate a process so that you can process more data. Data and statistics can also be used to make a case for a decision, or for critiquing an experiment or theory. Data and statistics are powerful types of information that can be used in myriad of ways.

Let us first provide some definitions on what we mean by data, statistics, aggregate data, and microdata. In this presentation, we are referring to data as a collection of facts, such as numbers, words, measurements, and/or observations. Most data we interact with in the 21st century is organized into numeric files created for the purpose of analysis and are machine readable. Keep in mind that data can be found in print documents as well.

Data can include geographic information, which can help our understanding of spatial patterns, especially when it is mapped. Lastly, data can include temporal information, which can be used to determine temporal patterns.

Statistics on the other hand are a type of information obtained through mathematical operations on data. Published statistics are processed datasets, or data that has been analyzed in some way. Statistics are used to support arguments or positions in studies, reports, and research articles. Many people will use the terms “data” and “statistics” interchangeably, because the definitions are so similar.

Microdata are un-summarized data or raw data. An example of microdata that we interact with daily is weather data. Each record in a microdata set represents a unit. In the case of weather data, it is a weather station.

In the social sciences and health sciences context, the unit is usually a person, but it can be a business, institution, or a household. Microdata provides information about the characteristics of the unit, such as the gender, age, income, and level of education of an individual.

Aggregate data, on the other hand, is created by combining unit level data. Unit level data are collected and combined from multiple units to create a statistical table that makes inferences about the population. Aggregated data provides high-level information about the population being studied, such as: after tax income of 25- to 34-year-olds in Toronto.

Now, let us focus a little more on microdata. Large microdata files are produced by many governmental agencies, and the largest government agency in Canada that produces microdata is Statistics Canada. In the context of Statistics Canada, they publish two distinct types of microdata:

A master file, which contains the original raw data, with each record denoting a person or an entity, with variables associated with that entity. Access to master data files is restricted because it contains sensitive and confidential information.

Public Use Microdata Files have records that are sampled from the master file, and the data has been anonymized and validated, so that it can be used for research purposes, but cannot be traced back to a specific individual.

This is an excerpt from the public use microdata for the 2016 Census of Population. On the left is how the file looks when you download it, and on the right is how it looks when you import it into a spreadsheet program such as Excel. You can use a microdata set to run statistical analysis and/or create your own aggregated dataset.

This is an example of an aggregate dataset, in which shelter costs are cross tabulated with the type of housing, in this case tenure. This data can then be transformed into charts to help communicate this information better, and help the reader identify patterns in the data.

The term, “disaggregated data” is quite common in the social sciences and health sciences, especially as it applies to race-based data, or other identity variables. Disaggregated data is broken down into component parts or smaller units of data for statistical analysis.

For example, in the context of race-based data, this means breaking down the composite (aggregate) “visible minority” or “racialized” category into its component parts such as Black, South Asian, East Asian, Southeast Asian, Latino, Middle Eastern, White, and others.

In the 2021 federal budget, funding was allocated to Statistics Canada to implement the Disaggregated Data Action Plan. According to Statistics Canada, “This funding will support more representative data collection methods, enhance statistics on diverse populations to allow for intersectional analyses, and support government and societal efforts to address known inequalities by bringing considerations of fairness and inclusion into decision-making.”

Let’s now focus on data collection methods. In the social sciences and health sciences, there are two data collection techniques that are common. One is a census, and the other is a survey. A census collects information about every member of a population and has a lot of detailed and accurate information. A census is an expensive undertaking, and it takes a lot of time for statisticians to analyze and release data from a census. The most common census is a national census, which most countries conduct. In Canada, it’s called the Census of Population and is conducted every five years.

On the other hand, a survey collects information from a sample of the population. Conducting and analyzing a survey is cheaper and less time consuming than a census, but the drawback is that a survey isn't as reliable or as accurate as a census.

The Census of Population is conducted by Statistics Canada. The Census of Population is a statistical portrait of Canada and Canadians on one specific day, taken every five years. Each housing unit or their household in Canada is asked to fill out the census, which collects various information, such as demographic, social, and economic characteristics.

These examples here provide you with an idea of the kind of sets of questionnaires in the census. The most recent census happened in May 2021, and the one before it in 2016. In May 2021, 100% of the households in Canada received a short form questionnaire, and it asked for your address, the date of birth of every individual living in the household, the gender or sex of every individual, the relationships of household members to each other, the knowledge of official languages and language you will speak at home, other languages that are spoken regularly, and your mother tongue.

Now, the long form census questionnaire goes to 25% of households. The long form asks the same questions as the short form questionnaire, but then it asks more detailed questions about ethnicity, visible minorities (racialized populations), indigenous peoples, education, housing, immigration, income, labor, how you get to work, knowledge of languages, mobility, and more.

And this is the data that the government uses to make policy. It's what researchers in academia use to figure out patterns in the Canadian population for research purposes, and this is going to be the primary source for demographic data related to ethnic groups, racialized persons and groups, and Indigenous peoples in Canada.

Keep in mind, it takes about a year for them to do all the number crunching and start delivering this information into different census data products. Most of the aggregate tables for the 2021 Census have been released, but we are still waiting for some specialized aggregate tables, the public use microdata files, and the master files to be made available.

A third common data collection technique is Administrative Data. Administrative data are information collected by government or private sector organizations as part of their ongoing operations. This type of data includes hospital discharge databases, birth and death databases, income tax databases, immigration databases, and more.

Module 2 Transcript

Welcome back to our Data Discovery series. You are watching Module 2: Search Strategy.

In this module, we will focus on understanding how to use a search strategy for effective data finding.

Similar to the process of effectively searching for literature, having a search strategy to find data is important. The four steps of the search strategy for data are as follows: First, define your area of study and research question. Second, determine the data producers of your topic or discipline and explore their resources. Next, narrow down resources and evaluate how well they meet your research question. Lastly, critique the dataset and producers.

In step one, we want to define our area of study and our research question, just like we do when searching for traditional literature. We want to define our concept keywords. We also want to think about the time period, geography, and any specific variables and data format we need or prefer.

In step 2, we want to determine what data is out there. It is important to explore the data resources that exist quite early in your research process. Sometimes, the exact data you need might not exist, particularly if it is a new or niche area of study. It costs money to collect, preserve, and distribute data

Thankfully, many times there is existing and relevant data, and it is a matter of searching and finding it. A good rule of thumb is to think about who would be interested in producing and sharing that data. For example, there may be governing bodies, non-profit organizations, or researchers who might be producing data on your topic and discipline

Module 3 will go more in-depth on how to determine data producers and other starting points for exploration

In step 3, you can narrow down your resources and evaluate how well they meet your research question. There might not be that special "unicorn" dataset, the more specialized and complex you go. We will explain this step in more detail in module 4

When you do find a dataset that’s of interest to you, it’s always important to critique and assess the dataset. It's important to understand the purpose of the data, determine if there are biases in the data, and answer pertinent questions related to methodology before you use it in your own research.

Module 5 will provide more details on how to critique data products

The search strategy is a naturally cyclical process. At any point in your searching, you might find yourself wanting to repeat it, determine new data producers, or find other resources, and that is completely normal. There might be a time when you have exhausted your knowledge of data producers, and at this point, you might want to reach out to York University Libraries' Data Services team for additional support.

Module 3 Transcript

Welcome back to Data Discovery. You are watching Module 3: exploratory research.

In this module, you will learn how to find and use the library research guide on data & statistics, and understand how to start finding major data producers.

As in module 2, let us use an example to illustrate a process to explore major data sources in this early stage of research. Our research question is "What is the impact of the housing costs on racialized populations in urban areas in Ontario?" with a couple of ideas for quantifiable data.

During the exploratory stage, we want to identify the broad topic first, so as not to rule out potential data producers too early. In this example, we can broaden to the topic of "Housing in Ontario" to investigate the resources that are available.

A good way to start understanding what information is available for a certain subject is to use the library's research guides. We can find the research guides from the library's home page by accessing the "Research & Learn" section, and under that, a "Research Guides" link. You can also access Research Guides with the button below the main search bar.

You can see that our research guides are grouped by subject area. Many might be of use to you, depending on your area of study. Today, we are going to click on the "Finding Types of Information" category, and then "Data & Statistics." This will take us to the home page of the "Data & Statistics" research guide. This guide is curated by the Data Services Team, past and present, and is based on topics that are commonly asked about.

This "Data & Statistics" guide contains many resources, from pages for different topic areas like the environment to human rights, as well as by different geographic areas. This home page also contains many quick links and descriptions to important data sources that are often quite general purpose and can be extremely useful for you. From the main page, we can also access more information about different types of data and how to use and cite data.

Today, we want to look at the topic guide dedicated to Housing that can be easily found under Topics (A-M) or from the drop-down menu.

The Housing topic guide is split into Canadian and Foreign/International. Since we're looking at affordable housing in Toronto, we'll focus on the Canada side. However, if we wanted to look into the international data that exists to compare the affordable housing in Toronto to other cities across the globe, we could also start looking at the Foreign/International side.

This section of the guide and today's exploration primarily focus on government data at all levels because the government creates policies on housing and thus needs to have data to decide how to formulate these policies. Governments then make some of this data available to the general public, students, and researchers.

We have Statistics Canada at the federal level that publishes tables, statistics, and analyses. They have recently started the Canadian Housing Survey in 2018, a survey conducted every two years dedicated to housing. In addition, the Census of Population has questions about

housing. We'll shortly see even more products. In this exploration, we are focused on Toronto housing data specifically and can see that the guide links to a few different City of Toronto sources on affordable housing.

Then, there is also the Canada Mortgage and Housing Corporation (CMHC), a crown corporation with the mission to make housing affordable. The CMHC collects data, analyses it, and publishes its data, along with research and reports.

Statistics Canada publishes microdata, aggregated datasets, infographics, visualizations, and more. You can access the Statistics Canada Housing Portal directly through the Housing topic page in our research guide. Or, you can also access it through the Statistics Canada website.

At the top of the page, click on Subjects, and select Housing. Here, we can find more than 100 related data products from 15 different survey or statistical programs. Let's click into the Housing Costs and Affordability subject area. Here, we find that there are 87 data sets and 27 analyses--the latter are reports or infographics. These analyses can be a helpful start for understanding what the government thinks is important from the data they're collecting.

Let's take a closer look at an individual dataset. With federal information, it's important to note the geography – we want data for Toronto, after all. Some datasets are only available at a larger level of geography, like province or territory, and this might not be as granular as we need. On Statistics Canada, the filtering is slightly hidden, but you just need to press the Add/Remove data button. Here, you'll see that we can narrow our selection to Toronto.

Other helpful parts of the tables are the citation instructions and DOI to the data. DOI stands for Digital Object Identifier, which is persistent and standardized. There is also the Related Information section, which shows us the source of the data, related products, and subjects to help with an organic exploratory search. In this case, I see that this data is from the Canadian Housing Survey, and I am curious if I can find out more about this survey.

Clicking on the Canadian Housing Survey, I am brought to detailed information on this specific survey. There is also a link to its questionnaire to know the exact questions on the survey. However, there is no microdata. How do we then access this microdata?

We can try ODESI to see if we can access these survey responses on an individual level. ODESI is digital repository for social science data, including polling data, with both microdata and aggregate data, as well as robust metadata.

We can use the quick links in the Data & Statistics research guide to access ODESI with our Passport York credentials. In ODESI, we can click Browse to see all of the available information by category. We can go to Social Surveys, then CANADA to find the Canadian Housing Survey. We can use ODESI to understand variables and survey responses for a specific variable, summarized. We can also use the Tabulation tool to create cross-tabulations.

ODESI allows us to download the whole dataset of microdata in a desired data format like SPSS or CSV, but it also provides a wealth of metadata information and data documentation. It is also a powerful tool to allow us to create tabulations and analyses right in the tool, like selecting a variable and immediately seeing its frequency.

The Housing section of the Data & Statistics guide also links to the Census of Population data on housing for the 2021, 2016, 2006, and 2001 censuses. Because the census goes to every Canadian household, it is filled with high-quality information on socio-economic trends, with many aggregated statistics, reports, and analysis tools built upon it by Statistics Canada, academic and private-sector researchers, and more.

Thus, there is a huge wealth of information, and to highlight just a few of the tables based on Census data, this screenshot on the right shows Statistics Canada products like a table on core housing need over time, a report on housing for Indigenous people in Canadian cities, and much more.

The next resource in our research guide is the Canada Mortgage and Housing Corporation data.

The CMHC collects data on core housing needs, the rental market, housing markets and more--all available in one place. Exploring this resource provides us with an understanding of what data variables and topics are collected on housing by the CMHC, as well as actual data. The CMHC also publishes several reports and research on housing in Canada.

Also linked within the Housing section of our research guide are City of Toronto data resources, since York University is situated within the GTA.

The City of Toronto considers affordable housing as a civic issue, with 52 datasets in its open data catalogue. Again, it is important to know what is available to us here before deciding on a research question, since it can help you then narrow down to a more specific topic—for example, maybe it is shelters you care about, or the amenities within apartment buildings.

Finally, linked inside the research guide but outside of the City of Toronto's Open Data portal, lie many reports and dashboards based on the open data, like this on the right about the shelter system flow.

Going to all of the individual open data portals of Ontario cities can take a while. We can also try using research data repositories, which allow data deposits from all levels of government and researchers at many academic institutions to be searched for and discovered. The Federated Research Data Repository, or FRDR, is affiliated with many Canadian institutions, and particularly governments, but there are many more research data repositories as well.

The FRDR is linked in the Quick Links section of the library's research guide. Once on the FRDR website, we can search "affordable housing" and see that we have thousands of results from organizations across Canada, including the City of Toronto and Ontario government datasets.

At York University, we subscribe to data from the Inter-University Consortium for Political and Social Research, and you can access it through the Quick Links section in the guide. ICPSR has a data archive of more than 250,000 files of research and data in social and behavioural science. The focus is predominately on the United States, but it also contains data sets that cover Europe, Latin America, Africa, and Canada. And, it's important to note that there are over 200 datasets just about Canada.

Another general purpose tool, and one that is comprehensive and multidisciplinary, is Statista. Again, you can access the full version through the Quick Links section in the guide.

Statista consolidates data from across thousands of sources, as well as using that data to create reports. You can search for a topic like “housing toronto”, and then find and have that data visualized for you, and importantly, also view what the source of the information is.

Furthermore, in our example, while we care about housing affordability in Ontario, we might want to understand affordability at a larger and international level. After all, housing affordability is a global crisis and many intergovernmental organizations and NGOs, from the OECD to the IMF, care about affordable housing. Both of these resources are listed in the Foreign/International tab within our Housing guide.

Module 4 Transcript

Welcome back to Data Discovery. You are watching Module 4: Evaluating data.

In this module, our main focus is on understanding how to evaluate data on a specific topic.

When we narrow down to a specific research question, there are key factors to consider when searching for and assessing specific datasets and products. These factors include: the data producer, geography, time period, level of observation, variables, data format, and the cost & access associated with the data.

As before, let us use our example topic. Our research question is "What is the impact of the housing costs on racialized populations in urban areas in Ontario?" with a couple of ideas for quantifiable data. More specific research questions include "How much income is spent on rent?" and "How much has it changed in the past 10 years?"

We must also ask the question: “Who is responsible for the collecting and publishing data about affordable housing?” Each level of government has different responsibilities when it comes to housing, in this case affordable housing. It is closely tied to legislation and policy as well. The major source at the federal level will be from Statistics Canada, as they collect data about income and census data, and conduct the Canadian Housing Survey. The provinces and municipalities also collect data about affordable housing. Other sources could include academic researchers, research institutes, non-profits, and lobby groups. It is also useful to list diverse ways the data are collected.

Before you start looking for a dataset, it is important to consider the level of observation and the population you are studying. In this case it is individuals from a racialized group. When you find the dataset, you must ask: Does it collect data about the population you are studying? And if so, how is that population defined and categorized?

Geography is also important. If we go back to our topic, our research focuses on "urban areas in Ontario". Depending on the dataset, we may get data at the level of Canada or the provinces/territories, but we may also get it down to the level of a census metropolitan area in Ontario, or a municipality in Ontario.

You should also consider the data's time period and frequency. It is important to ask these questions: What time period are you researching? Are you looking to compare data through time, and if so, with what frequency or intervals are the data available? You should also find out what is the most current data available, and what delays are there in the release of the data?

Now we want to think about what variables you may want to study or research. This provides you with clarity on what is being measured in the dataset that you find. It is also important to think of different search terms that you can use as this will assist you in understanding how a variable may be coded or categorized. This is where you examine the intersectional aspects of your topic, by including other demographic variables such as age, gender, immigrant/non-immigrant, and others. This will form your personalized checklist, which will help you compare different data products.

As for the format of the data source, this is also something you need to consider. Data are published in many different formats, and how you interact may depend upon your knowledge and ability of how to interpret and analyze data in a given format. Some data formats require a learning curve, such as data in a statistical software format, where you need additional technical skills and knowledge to manipulate and process the data in a statistical application. Whereas others may be easier to use, such as an aggregate data table or a chart.

Finally, we must consider the level of access and the data production and distribution cost. A lot of government information is free of cost and is available through open data portals, but even then, it is important to read the license agreements for data from open sites because the end-user agreements or license agreements will tell you how you can reuse or publish the data, as well as how to cite the data. This would be the case for most of the data on this research topic.

At York University, we are part of the Data Liberation Initiative. As a result, we subscribe to have access to the Public Use Microdata files from Statistics Canada, and you have access as a York University researcher via ODESI, but once you graduate you would have to contact Statistics Canada directly to have access to these files (or pay for access).

Other government data is kept confidential for security reasons or to protect the privacy of individuals or corporations. For example, when you fill out the census questionnaire, the information about you is private and confidential. In Canada, a researcher can go to a Statistics Canada Research Data Centre to obtain access to a master microdata file, in which only you as a researcher would have access to the data you request.

Since it is a long and rigorous vetting process, it is only recommended for faculty researchers or PhD candidates. When you are critiquing a dataset, look for documentation about the dataset. This can be a questionnaire, user guide, data dictionary, read-me file, or simply information on a website. If you can't find the information, you may have to contact a data librarian or the data producer for more information.

As a researcher there are some survey datasets in which you have access to description and documentation about the dataset, but you would have to apply directly to the data distributor to gain access via a mediated request. This is the case for data produced by the Canadian Institute for Health Information.

Finally, it is costly to produce data, and some data producers charge money to gain access to their data, this is especially true of data produced for commercial purposes. Governments also charge for access to their data. York University Libraries pays annual subscriptions to provide you with access to a variety of data sources and platforms, including ODESI, Statista, ICPSR (Inter-University Consortium for Political and Social Research), DMTI, SimplyAnalytics, and other major sources of data. So, if you ever see a dollar sign associated with a dataset you are interested in, contact us at datalib@yorku.ca to see if we have access to it already, or to find out if we have similar data that you can use.

With respect to aggregate data, this dataset can be found on Statistics Canada's data site. As you can see, this table is from the 2021 census. You can find data about visible minorities by

groups, the smallest level of geography is a census metropolitan area. We can get data about shelter costs to income ratio, core housing needs, age, gender, and immigration status.

What we can't get information about is if the housing is a rental or owned, the cost of the housing, income, or data at the level of municipality. Nor are data tables available for 2011 or 2016.

On the other hand, we can use the public use microdata file for the 2016 census, via ODESI. Like the previous table, we can get disaggregated data about visible minorities. The lowest level of geography is at the provincial level. We have a variable about tenure (rent/owned), income, age, and gender. Some, not all, metropolitan areas are included. The 2021 PUMF data hasn't been released yet. As for limitations, we can't get data at the municipality level, nor can we get the variable shelter cost to income.

Now let's move to the Canadian Housing Survey. When searching for data tables, I didn't find any on my topic. Instead, you can go to the Analysis portal on the Statistics Canada website to find reports, articles, and periodicals. Here I found several documents, including Housing experiences in Canada reports for 2021, which examine racialized groups, including Black, Chinese, and South Asian populations. In these documents, you will find data about monthly shelter costs, core housing needs, and dwelling conditions, but only at the federal level.

To get around these limitations, you can use the Canadian Housing Survey public use microdata file. There you will be able to get data down to the level of geography, with many of the variables on the checklist, but the visible minority variable is a yes or no variable. It isn't disaggregated. So, this gives you a sense of the data landscape for your research question. There isn't a one-size fits all dataset that answers the specific question.

Module 5 Transcript

Welcome back to Data Discovery. You are watching module 5: Critiquing data products.

In this module, you will learn to understand why it is important to critique a data product, and to be able to evaluate a found data product by four factors.

It is important to evaluate data and statistics for several reasons. It's crucial to understand how the data were compiled and constructed if you are going to be using it in your own research. You also want to know the context that the data were created in; it costs time, money, and human labour to create data, so it’s a good idea to know why the producers of the data wanted to do so in the first place.

After all, data and statistics, just like any other type of information, can lie. Although numbers may seem like uncontestable facts, they are just as easily manipulated or biased as any other type of text or visual out there.

Finally, we want to verify that the found data product fits your research needs, particularly that the currency, geographic coverage, and variables are aligned with your goals.

There are four factors, each with a set of guiding questions, for which you want to evaluate a data product. They are: Source and authority, objectivity and purpose, currency and geographic coverage, and methodology and completeness. Let’s explore each factor one-by-one in more detail.

First, you want to evaluate the source and authority of the data product. Who collected, produced, and/or published the data product? Keep in mind that the collector, producer, and publisher can be different entities.

If the entity collecting, producing, or publishing is an individual, it is important to determine who the individual is and how they are a reputable source for the topic. Is the individual affiliated with an organization, institution, or community, such as a researcher part of a policy hub or an NGO? If the entity is an organization, whether that be an agency, institution or community, then it is important to learn more about the entity and why they care about the data product.

It is also important to evaluate all of the entities related to the data product through its life cycle. If the data you found was repackaged or republished, it is best to determine the original source of the data to verify that the data were not negatively altered in the republishing and to determine the validity of the original source.

The second facet to evaluate is the objectivity of the dataset and your purpose for using it. All data products have intended purposes and audiences for the data, whether that be for the internal entity or for the wider community. For example, organizations ranging from York University to the Toronto Public Library might have administrative data that they use for internal purposes and decision-making, but they might decide to publicly share that data as well. Because there is always a cost associated with data, it is also salient to know who sponsored the data production and if there are any biases within the data product.

Related to objectivity, it is paramount to reflect on your own purpose and intent of using the data set, as well as to make sure you have noted how to properly cite the product.

The third facet to consider is the currency and geographic coverage of the data product. Particularly tied to how you intend to use it, you want to note the year the data was collected, when it was released or published, and how often it gets updated, if ever. You'll also want to take note of the geographical areas and the levels of geography that the data product covers.

Finally, the last facet of evaluation is on the methodology and completeness of the data. It is crucial to know exactly how the data were collected, whether that be a survey, census, or administrative data, that may or may not be flawed or biased. It is important to consider the intended population of the data, the methods used to select and sample the population, and the response rate. Remember that data is not exempt from racism and marginalization. You'll need to consider the populations that were included or excluded, particularly if those populations are underrepresented or systematically marginalized.

Additionally, it is good to see if the data can be verified, either by yourself or by other data sources that provide similar information and results.

After critiquing and evaluating the data products, you might find that you have finished your data search strategy. However, please note that the search strategy is a naturally cyclical process, and while conducting your evaluation, you might want to revisit your research question or find other data producers, and that is very common.

Module 6 Transcript

Welcome back to Data Discovery. You are watching Module 6: Citing data.

In this module, you will learn to understand the importance of data citation, recognize citation elements, and be able to apply different citation styles to cite data.

Just like other sources, such as books and journal articles, it is necessary to explicitly cite the data that you use. You should not just be citing a publication that references a data set, but the dataset itself, especially if you are reflecting on this data in your paper or project.

There are many reasons you do want to cite data, like giving credit to the original data producer and avoiding plagiarism, that is, reasons of academic integrity and contributing back to the ecosystem of scholarly communication.

Collecting and sharing data has real costs associated with it, and by citing that you used the data, you are demonstrating the value of that dataset and the impact it has. You also want to be able to prove that your data and the argument or analysis based on it, is from a credible source and that you can back that up with proof for your assignment.

The exact format of how to cite your data will depend on the citation style you choose to use, but including all of these elements, especially the elements in bold, are crucial regardless of style. You can find a more robust guide on how to cite at this link within the Data & Statistics Guide, the Citing Data subpage.

Some style guides, such as APA, ASA, or Chicago, will provide a specification for citing data, but others don't, like Harvard style. You can adapt based on the logic of your style or make use of the DataCite standard.

In APA format, this is an example of how to cite a dataset. As you can see, the standard looks very similar to how you would cite a book or article, with the author, the date, the title, and the URL. The new elements are specifying the format as a data file and providing the data distributor, which is ODESI in this example.