Introduction to Finding Data for NAN 708

Learning Objectives:

- Students will identify three common types of data sources and their differentiating characteristics.
- Students will evaluate data sources in order to select ones most relevant to an informational need or given problem.
- Students will develop search strategies for finding data based on their information need.

You need to find a dataset to use in an assignment or in your research. Where do you start? In this tutorial, we'll look at common sources of data, where to find them, and how to identify keywords to use in your search.

Click "next" below to go to the next page in this module.

Sources of data

You can find secondary data, or data collected for a primary project or purpose and then shared for reuse in other, secondary projects, in many different types of places, both in print and online. Some data, especially from data producers within the open data community, is made available to all for free through websites like **GitHub (https://github.com/)**, **Wikidata (https://www.wikidata.org/wiki/Wikidata:Main_Page)**, or **OSF (https://osf.io/)**. Data funded or collected by government agencies, educational institutions, and news organizations, especially research data, can be found in a variety of places. There are a lot of places you can look!

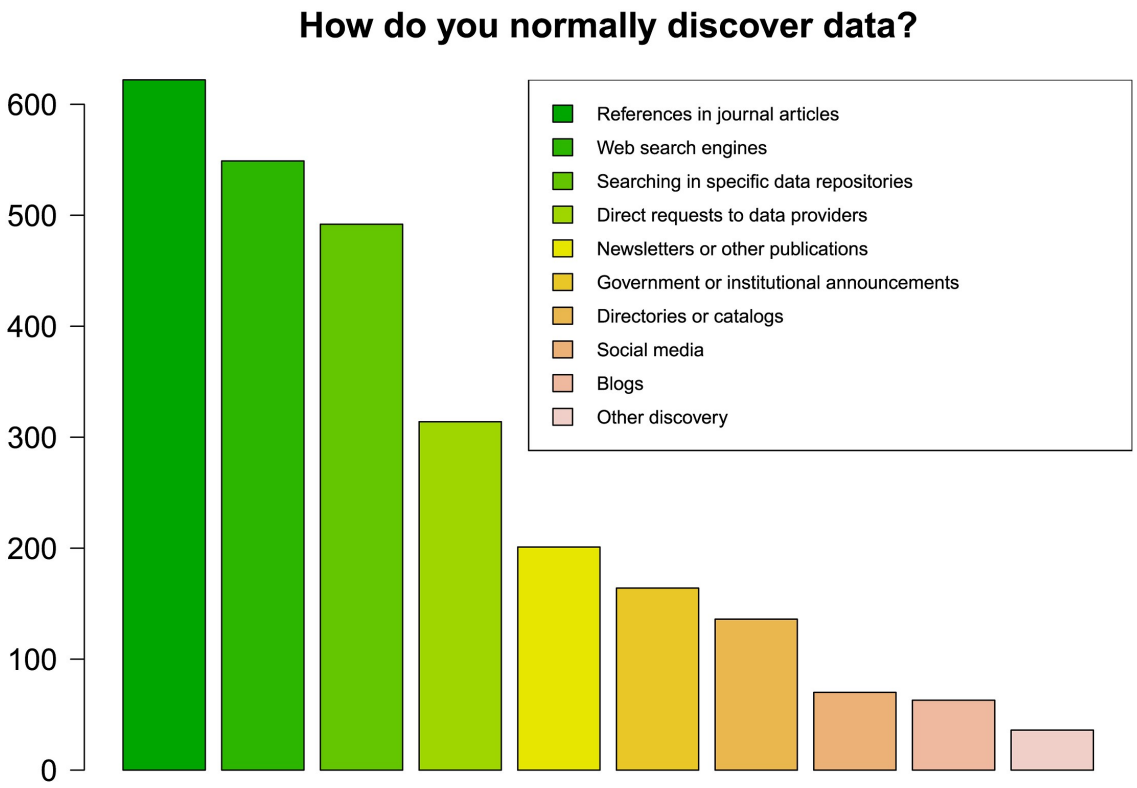


Image: “Fig 9: Discovery of data.” How 774 surveyed users and providers of open data in the scope of global environmental change normally discover data (Schmidt et al., 2016).

Knowing what type of place you’re looking in can help you choose a search strategy and give you an idea of what data you should expect to find there, as well as what tools and skills you’ll need to actually acquire and use the data.

Three common types of research data sources are:

- Data repositories
- Web-based data browsers, catalogs, and portals
- Books, visualizations, and other publications

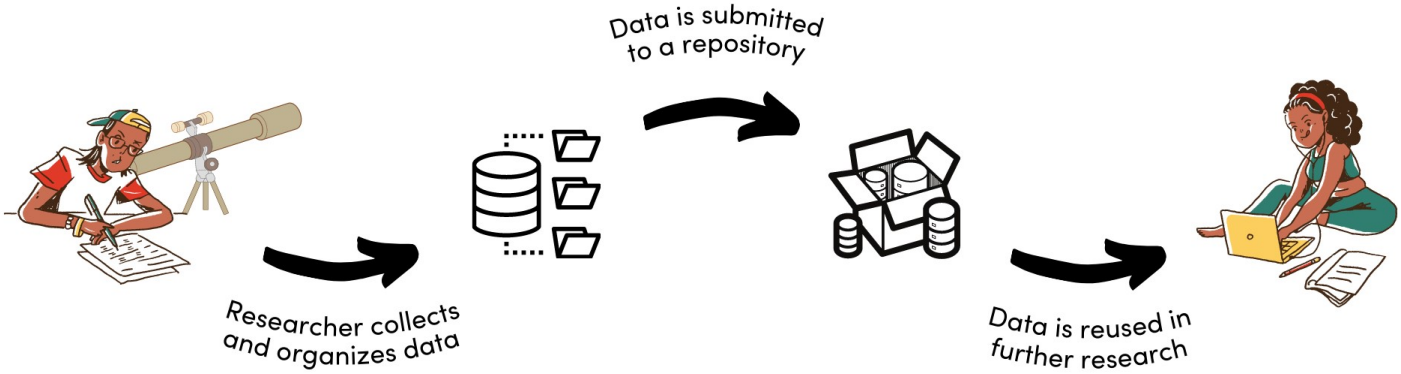
Click "next" below to go to the next page in this module.

Sources on this page:

Schmidt, B., Gemeinholzer, B., & Treloar, A. (2016). Open data in global environmental research: The Belmont Forum's open data survey. *PLOS ONE, 11*(1). **https://doi.org/10.1371/journal.pone.0146695 (https://doi.org/10.1371/journal.pone.0146695)**

Data repositories overview

The first type of common data source is a data repository. According to the Network of the National Library of Medicine (NNLM), "a data repository can be defined as a place that holds data, makes data available to use, and organizes data in a logical manner… [or] an appropriate, subject-specific location where researchers can submit their data" (n.d.). These repositories can be associated with physical places, or be entirely online. They can also be run by a variety of entities, including universities, government agencies, non-profits or groups of volunteers, or publishers and other for-profit companies. Some data repositories have a review process to vet data that is submitted, while others don't. Some data repositories are discipline specific and only accept data related to that area of study - these are a great way to narrow down your search. Depending on your area of study, you may have a data repository specifically for your discipline!



Many data repositories have restrictions on who can deposit data into them based on funding, academic qualification, and quality of data. Some are *open*, meaning almost anyone can add, or *deposit*, data. Don't confuse this with *open access* or *open data* repositories, which allow anyone to access, use, and share the data, but may still limit who can deposit data (Open Knowledge Foundation, 2015). You can find more information and examples of repositories at:

- **NNLM Thesaurus | Data Repository (https://nnlm.gov/data/thesaurus/data-repository)**
- **Registry of Research Data Repositories (re3data.org) (https://www.re3data.org/)**
- **Open Access Directory (OAD) Repository List (http://oad.simmons.edu/oadwiki/Data_repositories)**

The UNCG Libraries provide access to some data repositories and sources through subscriptions or institutional memberships. You may see these repositories, and the collections of books, journals, and other library materials that the Libraries also provide access to (for example, JSTOR and Science Direct) referred to as *library databases*.

Quick Check:

Thank you for trying out H5P. To get started with H5P read our getting started guide

There may be a data repository for your specific discipline.

○ True ○ False

Check

Reuse <> Embed H5P

Sources on this page:

Network of the National Library of Medicine. (n.d.). *Data repository.* National Library of Medicine, National Institutes of Health. Retrieved March 8, 2021, from https://nnlm.gov/data/thesaurus/data-repository (https://nnlm.gov/data/thesaurus/data-repository)

Open Knowledge Foundation. (2015, August 18). *What is open data?* http://opendatahandbook.org/guide/en/what-is-open-data/ (http://opendatahandbook.org/guide/en/what-is-open-data/)

Types of data repositories

Closed, restricted, or limited access repositories

Some repositories, like the **Inter-University Consortium for Political and Social Research (ICPSR) (https://library.uncg.edu/dbs/auth/go.aspx?vdbID=606)** , limit access of some or all data to member institutions only, or only share *metadata* - data about the data, like creator, title, or date collected - through citations, and require you to request to download and use some datasets. This can help researchers control who is able to access their shared data, especially if it contains sensitive or confidential information or information which could put subjects of the data at risk.

Data warehouses and clearinghouses

Larger repositories that collect, organize and make available data from multiple sources (including other repositories) are often referred to as data *warehouses* or sometimes *clearinghouses*. Some are open access, and others provide access through subscriptions and offer advanced features like analysis tools and specialized datasets for subscribers.

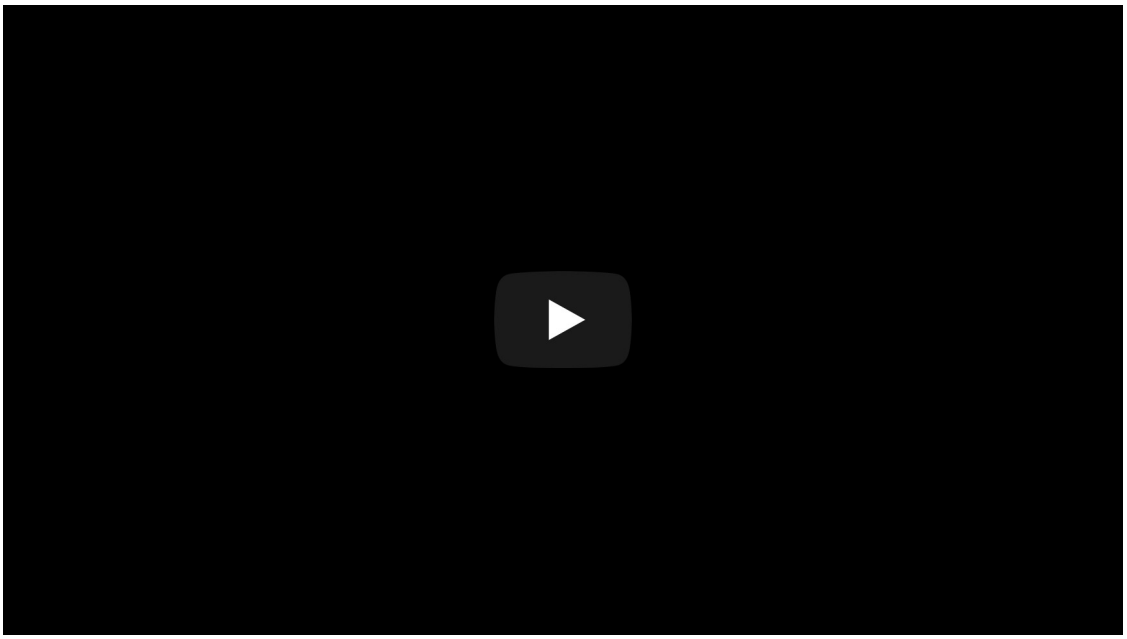
- Some warehouses collect data within a specific theme; for example, **DataONE (https://www.dataone.org/)** provides access to data related to Earth and environmental science.
- Some are less like repositories as they don't allow for depositing data, and instead collect and repackage data themselves; for example, **PolicyMap (https://library.uncg.edu/dbs/auth/go.aspx?vdbID=1312)** has data relevant to public policy across multiple disciplines, but gets it from public and commercial sources rather than having researchers submit their data.

Institutional repositories

Institutional repositories collect data produced by researchers affiliated with the institution. Institutional repositories make it easier for those researchers to collaborate on projects and share their data with other researchers at that institution, especially if their data doesn't meet the requirements of other, often discipline-

specific repositories. For example, UNCG partners with UNC Chapel Hill's Odum Institute to host research data produced by UNCG faculty and other researchers in the **UNC Dataverse (https://dataverse.unc.edu/dataverse/UNCG)** . This data doesn't have to be about UNCG, and often isn't!

This video from the Odum Institute will tell you more about the UNC Dataverse and how it works:



Quick Check:

Thank you for trying out H5P. To get started with H5P read our getting started guide

You can use any data in the Inter-University Consortium for Political and Social Research (ICPSR) repository without getting special permission .

○ True ○ False

Check

Reuse <> Embed H5P

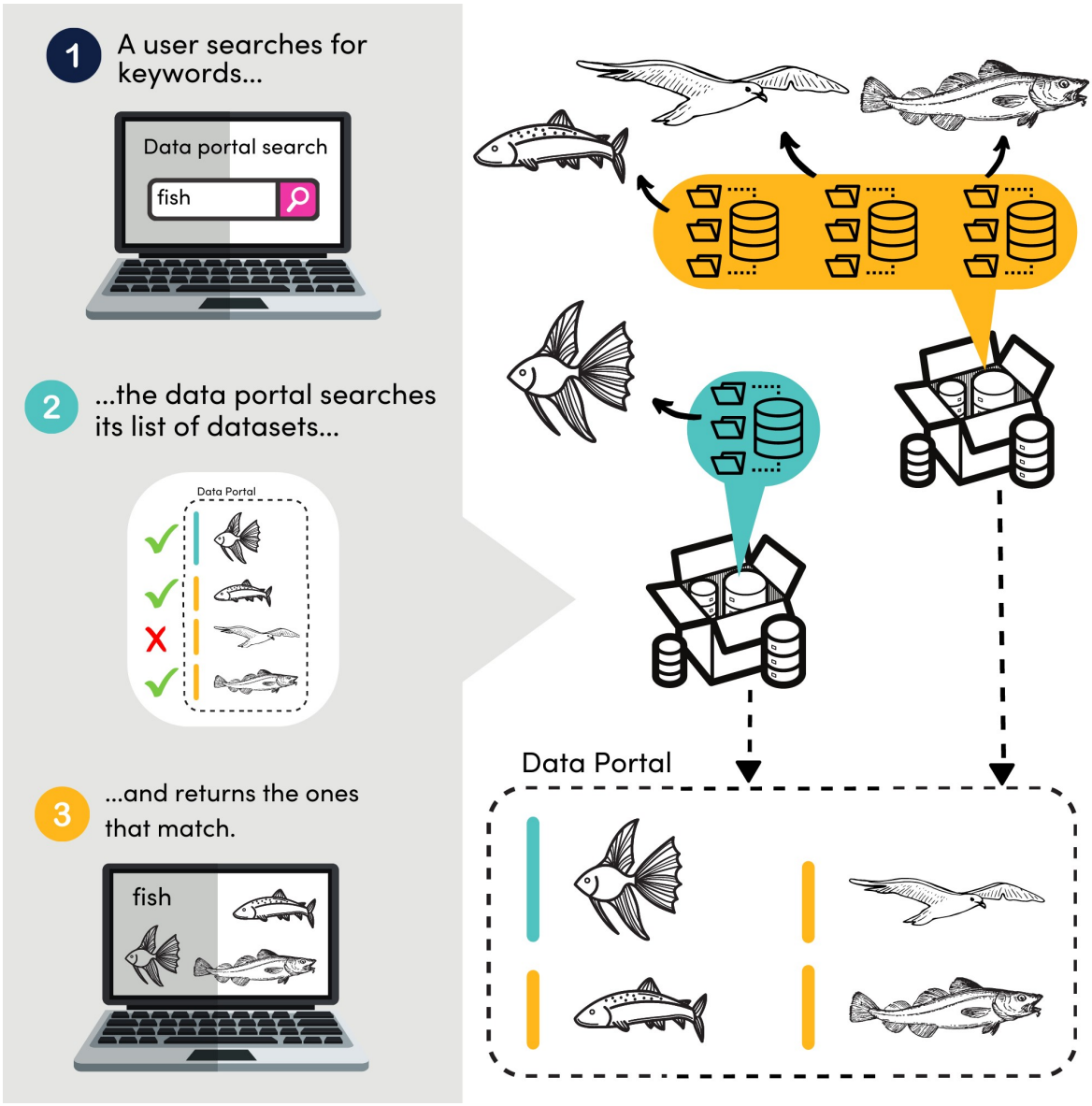
Sources on this page:

Odum Institute. (2017, January 11). *Dataverse - overview* [Video]. YouTube. **https://www.youtube.com/watch?v=4eSh9OaCqGU (https://www.youtube.com/watch?v=4eSh9OaCqGU)**

Web-based data browsers, catalogs, and portals

Web-based data browsers, sometimes called *portals* or *catalogs*, often work similarly to Google and other search engines: a user like you searches for keywords, the search engine looks for those keywords in its list of metadata, and then shows the user a smaller list of datasets that match those keywords.

These data browsers, like **Google Dataset Search (https://datasetsearch.research.google.com/)** , usually don't store the data and make it available, but rather compile citations for the data that link out to other websites and repositories where the data is stored and made available. You'll likely need to click through to another website in order to download the data and its associated documentation (e.g. methodology, terms of use, and other important information), and that data may not always be available to you or be a credible source of information depending on the source. This is similar to using Google Scholar to find scholarly articles.



Government entities, like the US Environmental Protection Agency, US Census Bureau, and some state and local governments, may have a data portal to help you find data produced by one or more government agencies, often across multiple repositories at once. These are primarily sources of *authoritative* data, from a legally authorized or officially recognized source. Some portals, especially those of local or state governments, have characteristics similar to data repositories and fit the technical definition of a repository, but others don't.

Some portals, like EnviroAtlas, provide access to their own data alongside data pulled in from other repositories and specialized analysis tools to help you interpret and combine data. These often include an interactive map or data visualization

feature and you'll sometimes see the entire website referred to as an app, webapp, or tool.

Larger datasets and data that are not available for download online sometimes may be findable in a data portal or browser, but will need to be sent to you on CDs, USB storage devices, or via an emailed link or file transfer methods such as FTP or API connection. FTP and other file transfer methods are especially common when working with *big data*. Be aware of this when looking for data in these types of places!

Quick Check:

Thank you for trying out H5P. To get started with H5P read our getting started guide

When using data portals you may need to click through several websites to get to the original source of the data

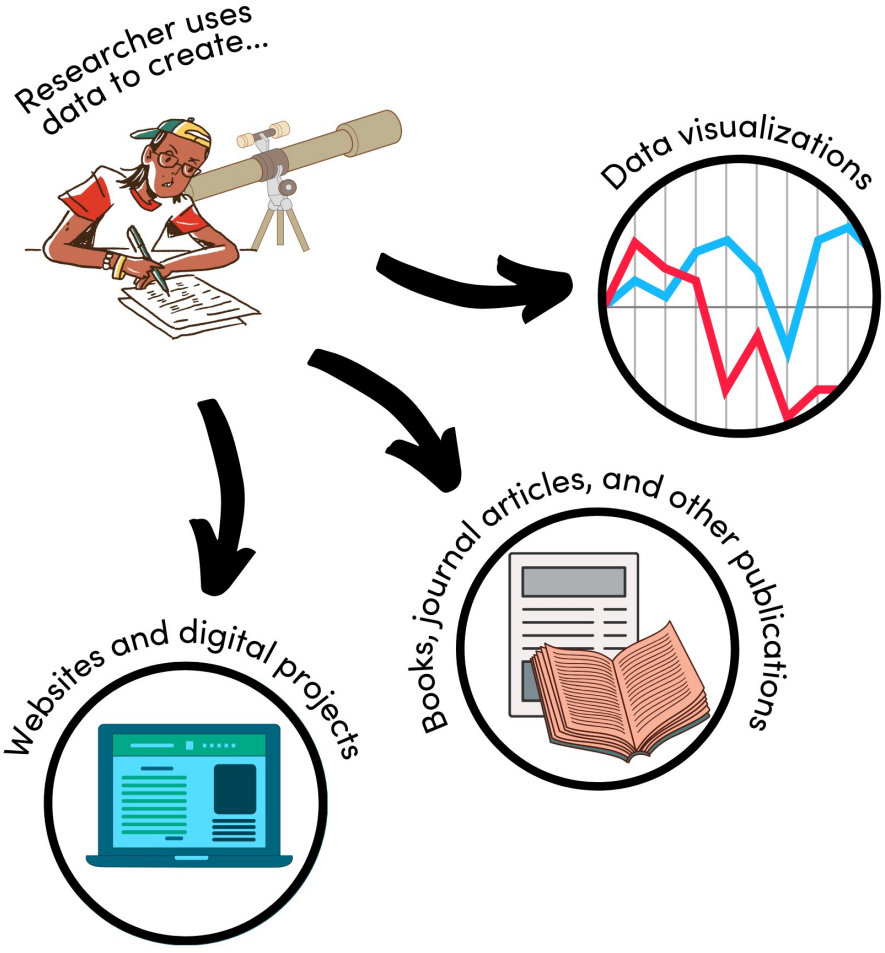
○ True ○ False

Check

Reuse <> Embed H5P

Books, visualizations, and other publications

Journal articles, books, reports, data visualizations like charts and infographics, and sometimes websites and other digital projects created using data - like the **New York Times COVID-19 Map and Case Count site (https://www.nytimes.com/interactive/2020/us/about-coronavirus-data-maps.html)** (**source on GitHub (https://github.com/nytimes/covid-19-data/blob/master/)**) - should cite the data or data collection method. The creators of data visualizations and digital project websites will sometimes directly link out to the data they used with a link that says “source” or “source data,” although source can also refer to the creator of the visualization or website itself.



It can take longer to find data using these sources, but it's a good way to find data related to a very specific topic or project. It's also a good way to start your search if you're not exactly sure what data will help you answer your research question; for example, an online mapping tool that you found using a Google search of your topic can give you an example of what data and keywords to look for, plus a trail of documentation to follow back to the original data collector and their other work.

Some data, especially historical datasets, are published as tables in factbooks, reports, books, or alongside other types of publications both digital and in print. You can find many of these ready for download or use online, but some will indicate that they're only available for in-person use at a library or archive. Data can also be published in the appendix or multiple appendices at the end of a book, or as a table or chart in figures interspersed throughout a publication. You can sometimes find these publications using the library catalog, but it's difficult to filter your search specifically for publications that contain data unless it is a factbook or report.

Quick Check:

Thank you for trying out H5P. To get started with H5P read our getting started guide

You can only find data online.

○ True ○ False

Check

Reuse <> Embed H5P

Summary: Sources of data

In summary, you will need to adapt your search strategy depending on the type of data you need to answer your research question, and where you might find that data.

Data repositories

Data repositories collect, organize, and provide usually direct access to data. They include institutional repositories, discipline-specific repositories, and thematic data warehouses or clearinghouses.

Web-based data browsers, catalogs, and portals

Data browsers, catalogs, and portals collect and provide access to information about data across repositories and other sources, often related to a theme or within a federal, state, or local government jurisdiction.

Books, visualizations, and other publications

Books, visualizations, and other publications created using data provide access to that data or to information about the data, related to a specific topic or research question.

Photo by Cookie the Pom (https://unsplash.com/@cookiethepom?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText) on Unsplash (https://unsplash.com/?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText)

Click "next" below to go to the next page in this module.

Finding Sources

So where do you find these sources?

Library Catalog

Use the library catalog to find publications that might contain data or that might cite a data source, keywords to help you develop your search strategy, and background information to help you interpret the data. You may find older data files in the catalog too, by filtering your search to “computer files.”

Library Databases

Use the **library databases list (https://uncg.libguides.com/az.php?t=37851)** to find themed data repositories and warehouses that you get access to through UNCG like PolicyMap, SAGE Data Planet, SimplyAnalytics, and ICPSR.

Research Guides

Use research guides to find links to library resources and websites, including library databases, repositories, and any other type of data source available to UNCG students and faculty, curated by a librarian for a specific subject or course. The **data guide (https://uncg.libguides.com/data/datasets)** lists frequently used data sources organized by popular topics, or you can find your course guide in Canvas or in the **guides list (https://uncg.libguides.com/)** .

Other Websites

Data sources, like other sources of information, can be found just about anywhere on the internet as well:

- Government data portals can usually be found on the website for the government jurisdiction (for example, Greensboro has an open data portal, **Open Gate City (https://data.greensboro-nc.gov/)** , that you can find on the **City of Greensboro's website (https://www.greensboro-nc.gov/)**), or you can search for “data” and a keyword or two on government websites to find related data.

- Google Dataset Search and other portals and repositories can be found using a search engine like Google, as can a variety of other websites and blogs that compile datasets for use in instruction and learning, like **Data is Plural** (https://tinyletter.com/data-is-plural) and **TidyTuesday** (https://github.com/rfordatascience/tidytuesday). Take care to evaluate data you find this way for credibility and suitability for your research needs - some of it will not be research-quality data.
- Visualizations and other online publications like digital project websites are generally findable using a search engine, if you know the right keywords to search for. Some aren't, and you'll need to do some reading to find background information and potential leads.

Photo by **Jan Antonin Kolar** (https://unsplash.com/@jankolar?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText) on **Unsplash** (https://unsplash.com/?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText)

Quick Check:

Thank you for trying out H5P. To get started with H5P read our getting started guide

(Select all that apply) You can find data repositories using:

- [ ] Library databases
- [ ] The library catalog
- [ ] Other websites
- [ ] Research guides

Check

Reuse <> Embed

Identifying keywords

Because there are so many different sources to look through, it can be tough to figure out how to start your search, and what to look for. As you get more familiar with these sources, it'll get easier to narrow your search down to the relevant ones and knock out irrelevant ones. One strategy to make this easier is to know what you're looking for - identify keywords to look for while browsing or to use in searches. If you've searched Google or have had to find sources for a research assignment, you're probably already familiar with the concept of *keywords* or search terms. If not, check out **Find: Creating Keywords. (http://libapps4.uncg.edu/tutorials/module.aspx?t=63&m=81)**

In addition to using keywords to search for data, you'll also use them to learn more about a dataset and whether it fits your needs. You can find keywords in the search filters of most data sources, in the description of a dataset and other data documentation, and using tools like the **ICPSR Variables Browser (https://www.icpsr.umich.edu/web/pages/ICPSR/ssvd/)** . Data documentation can take many forms depending on the preferences of the data creator and where they shared their data; look for words like codebook, index, data key, layout, metadata, description (or "about"), and methodology.

Photo by Mitchell Luo (https://unsplash.com/@mitchel3uo?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText) on Unsplash (https://unsplash.com/?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText)

One way to identify keywords to help you search for and learn more about data is by using the “5 Ws and an H” framework:

Who

Who is the source of the data, i.e. who collected and organized it or published it? There might be multiple people or groups involved with the creation of a dataset. If you’re looking for data produced by researchers at a university, you might use the university name as a search term in a data repository, or even search within that university’s institutional repository.

What

What is your data about, and what data characteristics are you looking for? Think about how variables might be worded and keep an eye out for synonyms, for example “train” vs “rail” vs “metro” in public transportation data. Also, think about what files and formats you can work with. Some repositories will have filters that you can use to narrow down your search to certain file formats or sizes.

When

When was the data produced, what time period does the data cover, and what else was happening around those times? For example, the Decennial Census is taken every 10 years - you won’t be able to find Decennial Census data for 2018, and will need a different dataset. Release of regular datasets might also be interrupted by events like the COVID-19 pandemic, which disrupted data collection due to travel and work restrictions. Remember that the date a dataset was published may not be the same as the date it was collected.

Where

What locations does the data cover, and if you’re looking for geographic data for use

in GIS, what coordinate reference system will you need to look for? You may be more likely to find data about a specific city from a state-level source rather than a national source, for example. Also consider where the data was collected and published. Some countries do not collect types of data that others do, or do not make it publicly available.

Why

Why was the data collected? *Administrative data*, collected by an organization as part of routine services, will likely be organized and shared differently than *research data* collected for a study. If data was collected for a specific survey, like the American Community Survey, you'll likely find it right on the American Community Survey website, and might find related information there too. Think about the likely audience of a published dataset too; if the audience is researchers within a specific discipline that you're not familiar with, there may be words or variables you need to look up or find background information on.

How

How was the data collected? If it's survey data, how were the questions worded? If specialized equipment was used, do you need the methodology and error rate, and can you find this information in the documentation for a dataset? Some data is estimated using models and calculations instead of collected through observation, which will change what keywords you use.

Quick Check:

Thank you for trying out H5P. To get started with H5P read our getting started guide

I am looking for streamflow measurements collected by Bart Greenjeans in 2010 along the Neuse River to compare with data from the USGS National Water Information System.

What keywords should I use to narrow down my search based on “Who”?

- ○ Streamflow measurements
- ○ Bart Greenjeans
- ○ USGS National Water Information System

Check

Reuse <> Embed H5P

NAN 708 Library Research Module Assessment

Thank you for completing the NAN 708 Library Research Module on Finding Data!

Please complete the form embedded below (**or you can open it in a separate window by clicking here (https://docs.google.com/forms/d/e/1FAIpQLSeJQDTxaS0KJBonsIVxotIYkciOysk7UcT0vIUAsBAIPM4FXA/viewform?usp=sf_link)**). This form is not a graded assignment or quiz, and it is set up to be anonymous. Your responses will help me assess how effective this module is and will help me prepare for Thursday's class.

NAN 708 Library Research Module: Finding Data Assessment

Answer these questions to the best of your ability.

What's one thing you learned in this module?

Your answer

What's one thing that's still confusing?

Your answer

What questions do you have about Thursday's class?

Your answer

Submit

Never submit passwords through Google Forms.

NAN 708 Library Research Module: Finding Data Assessment

4 responses

Publish analytics

What's one thing you learned in this module?

4 responses

- Where to find data for my research work
- About data repositories.
- How/What databases help us to look for data for our research
- How to use data repository and sources to find them.

What's one thing that's still confusing?

2 responses

- The exact pathway to access data in the library website
- How can people edit the data on the databases?

What questions do you have about Thursday's class?

1 response

- None