VANDERBILT UNIVERSITY

BME 7410

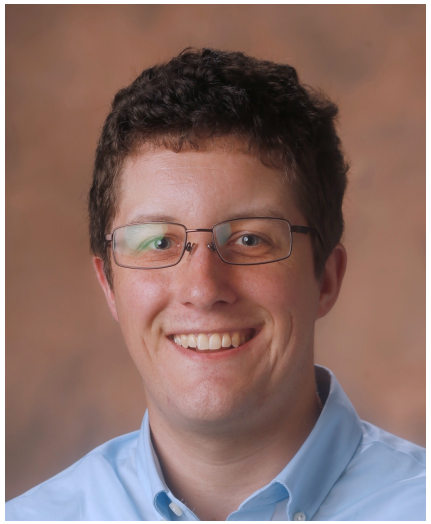
Alex Carroll, MSLS, AHIP
September 8, 2022

Created by David Khai
from the Noun Project

Our Goals for Today

- Share a framework for evaluating open data repositories
- Use that framework to evaluate a repository together
- Explore other options for discovering data sets

STEM Librarian Team

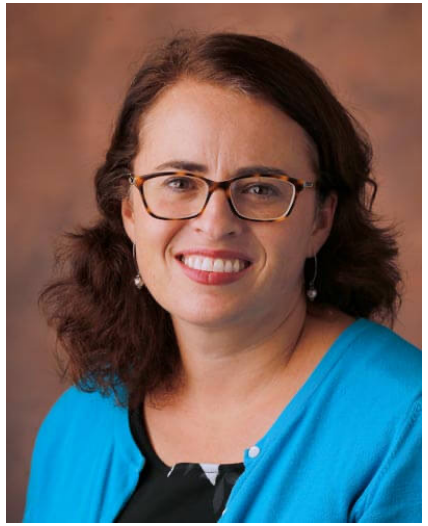


Josh Borycz
joshua.d.borycz@vanderbilt.edu

vanderbi.lt/borycz

Alex Carroll
alex.carroll@vanderbilt.edu

vanderbi.lt/carroll



Honora Eskridge
honora.eskridge@vanderbilt.edu

vanderbi.lt/eskridge

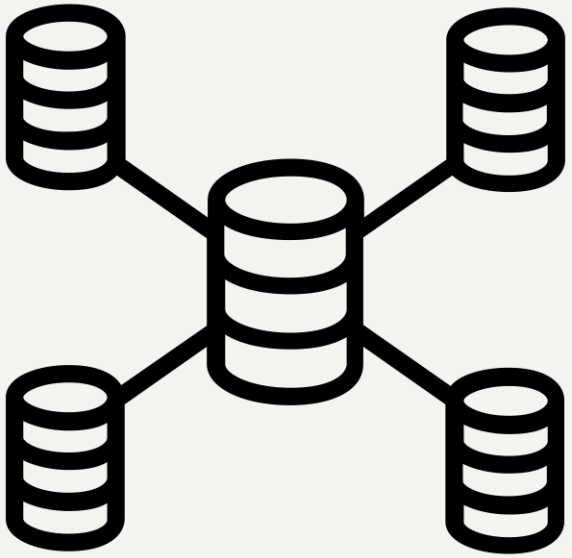
Francisco Juarez
francisco.juarez@vanderbilt.edu

vanderbi.lt/juarez

Evaluating Open Repositories

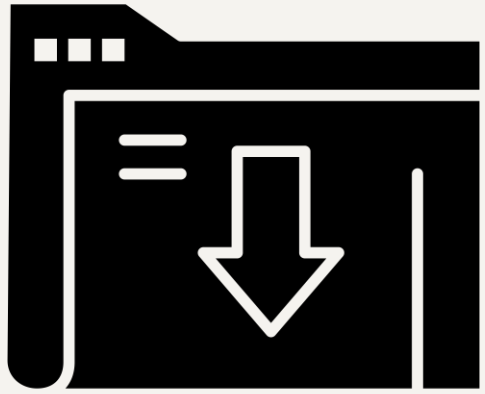
Evaluating data repositories is a two step process

Q1: What are the attributes of the repository?



Created by ProSymbols from Noun Project

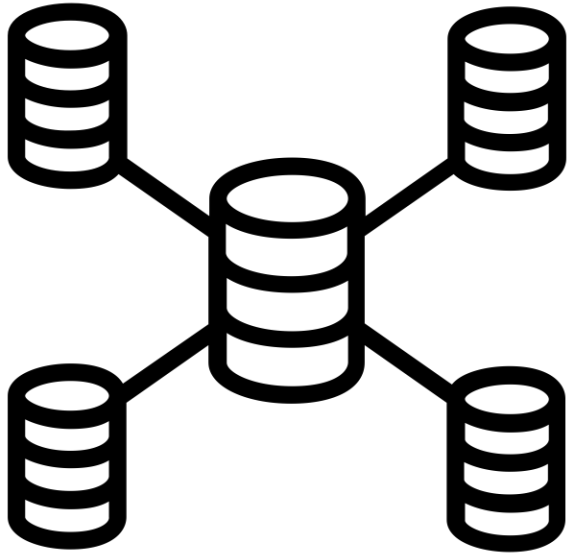
Q2: What are the attributes of the data stored in it?



Created by Flatart from Noun Project

Evaluating the repository

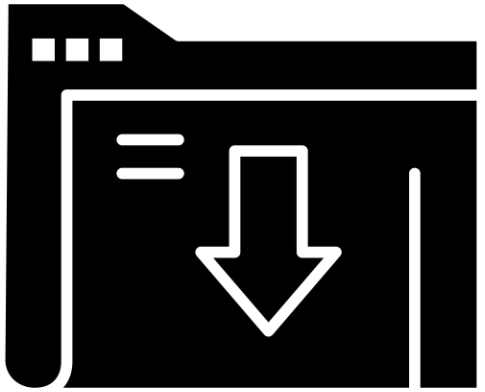
- Who manages this repository?
- When was the repository last updated?
- What sources are the data drawn from?
- Who can deposit data into the repository?
- How can you search for data within the repository?



Created by ProSymbols
from Noun Project

Evaluating the data

- What time scale of data is included?
  - 2010-2017, 1995-2015
- What documentation is available for these data?
  - Data dictionary, README
- What access restrictions exist?
  - Are permissions required? Are only certain kinds of use allowed?
- How can the data be accessed?
  - Zip file downloads, FTP, API, HTML only
- In what formats can data be exported?
  - HTML, tab-deliminated, txt



Created by Flatart
from Noun Project

Applying this framework

Who manages this repository?

United States Cancer Statistics

Public Information Data

United States Cancer Statistics Data

Current Cancer Statistics

- **Cancer Incidence 1999 - 2018**: By year, state, metropolitan area, age group, race, ethnicity, sex, childhood cancers and cancer site classifications.
  Data Request More information
- **Cancer Mortality 1999 - 2018**: By year, state, metropolitan area, age group, race, ethnicity, sex, and cancer site classifications. See below for more information on mortality rate comparisons.
  Data Request More information
- **Cancer Mortality Incidence Rate Ratios 1999 - 2018**: By year, state, metropolitan area, race, ethnicity, sex, and cancer site classifications. See below for more information on mortality rate comparisons.
  Data Request More information
- **National Program of Cancer Registries 5 year Relative Survival**: By race, sex, age group and cancer site classifications. Current NPCR 5-year survival statistics are available at United States Cancer Statistics: Data Visualizations.

Archive Cancer Statistics

CDC WONDER maintains archive versions of previous releases of data on this website to allow users to replicate data requests that were conducted in the past. Please refer to the main data repositories, listed above, to access current data.

Click here to access the archived data.

More Information

The United States Cancer Statistics (USCS) are the official federal statistics on cancer incidence from registries having high quality data and cancer mortality statistics for 50 states and the District of Columbia. USCS are produced by the Centers for Disease Control and Prevention (CDC) and the National Cancer Institute (NCI). For a list of all USCS contributors and partners, visit USCS Contributors.

Incidence data are provided by:

- The Centers for Disease Control and Prevention National Program of Cancer Registries (NPCR)
- The National Cancer Institute Surveillance, Epidemiology and End Results (SEER) program

the Centers for Disease Control and Prevention National Vital Statistics System (NVSS). Cancer mortality data for deaths after 2018 are available from NVSS.

- **Cancer Incidence 1999 - 2018**: By year, state, metropolitan area, age group, race, ethnicity, sex, childhood cancers and cancer site classifications.
  Data Request More information
- **Cancer Mortality 1999 - 2018**: By year, state, metropolitan area, age group, race, ethnicity, sex, and cancer site classifications. See below for more information on mortality rate comparisons.
  Data Request More information
- **Cancer Mortality Incidence Rate Ratios 1999 - 2018**: By year, state, metropolitan area, race, ethnicity, sex, and cancer site classifications. See below for more information on mortality rate comparisons.
  Data Request More information
- **National Program of Cancer Registries 5 year Relative Survival**: By race, sex, age group and cancer site classifications. Current NPCR 5-year survival statistics are available at United States Cancer Statistics: Data Visualizations.

Archive Cancer Statistics

CDC WONDER maintains archive versions of previous releases of data on this website to allow users to replicate data requests that were conducted in the past. Please refer to the main data repositories, listed above, to access current data.

Click here to access the archived data.

Who can deposit data into the repository?

More Information

The United States Cancer Statistics (USCS) are the official federal statistics on cancer incidence from registries having high-quality data and cancer mortality statistics for 50 states and the District of Columbia. USCS are produced by the Centers for Disease Control and Prevention (CDC) and the National Cancer Institute (NCI). For a list of all USCS contributors and partners, visit USCS Contributors.

Incidence data are provided by:

- The Centers for Disease Control and Prevention National Program of Cancer Registries (NPCR)
- The National Cancer Institute Surveillance, Epidemiology and End Results (SEER) program

the Centers for Disease Control and Prevention National Vital Statistics System (NVSS). Cancer mortality data for deaths after 2018 are available from NVSS.

About mortality rate comparisons:

- For consistency with the data on cancer incidence, the cancer sites in mortality data were grouped according to the revised SEER recodes dated March 1, 2018 (see SEER Cause of Death Recodes). Because NCHS uses different groupings for some sites, the death rates in this report may differ slightly from those published by NCHS.
- The population used to age-adjust the rates in this report is the 2000 U.S. standard population, which is in accordance with a 1998 recommendation of the US. Department of Health and Human Services. The 2000 U.S. standard population is based on the proportion of the 2000 population in specific age groups (younger than 1 year, 1-4 years, 5-9 years, 10-14 years, 15-19 years, . . . 85 years or older); the proportions of the 2000 population in these age groups serve as weights for calculating age-adjusted incidence and death rates. NCHS, however, uses a different set of age groups in its age adjustment of death rates, and thus the cancer death rates in this report may differ slightly from those published by NCHS.
- Deaths of persons of unknown age are not included in this data set. Death rates may differ slightly from other reports where deaths of persons of unknown age are included.

When was the repository last updated?

What sources are the data been drawn from?

Page last reviewed: October 07, 2021
Content source: CDC WONDER

CDC WONDER FAQs Help Contact Us WONDER Search

How can you search for data within the repository?

United States Cancer Statistics

Public Information Data

United States Cancer Statistics Data

Current Cancer Statistics

- **Cancer Incidence 1999 - 2018**: By year, state, metropolitan area, age group, race, ethnicity, sex, childhood cancers and cancer site classifications.
  Data Request More information
- **Cancer Mortality 1999 - 2018**: By year, state, metropolitan area, age group, race, ethnicity, sex, and cancer site classifications. See below for more information on mortality rate comparisons.
  Data Request More information
- **Cancer Mortality Incidence Rate Ratios 1999 - 2018**: By year, state, metropolitan area, race, ethnicity, sex, and cancer site classifications. See below for more information on mortality rate comparisons.
  Data Request More information
- **National Program of Cancer Registries 5 year Relative Survival**: By race, sex, age group and cancer site classifications. Current NPCR 5-year survival statistics are available at United States Cancer Statistics: Data Visualizations.

Archive Cancer Statistics

CDC WONDER maintains archive versions of previous releases of data on this website to allow users to replicate data requests that were conducted in the past. Please refer to the main data repositories, listed above, to access current data.

Click here to access the archived data.

More Information

The United States Cancer Statistics (USCS) are the official federal statistics on cancer incidence from registries having high-quality data and cancer mortality statistics for 50 states and the District of Columbia. USCS are produced by the Centers for Disease Control and Prevention (CDC) and the National Cancer Institute (NCI). For a list of all USCS contributors and partners, visit USCS Contributors.

Incidence data are provided by:

- The Centers for Disease Control and Prevention National Program of Cancer Registries (NPCR)
- The National Cancer Institute Surveillance, Epidemiology and End Results (SEER) program

the Centers for Disease Control and Prevention National Vital Statistics System (NVSS). Cancer mortality data for deaths after 2018 are available from NVSS.

United States and Puerto Rico Cancer Statistics, 1999-2018 Incidence Request

Request Form | Results | Map | Chart | About

Cancer Statistics Data | Dataset Documentation | Other Data Access | Data Use Restrictions | How to Use WONDER

Save | Reset

*Make all desired selections and then click any **Send** button one time to send your request.*

1. Organize table layout:

Send | Help

Group Results By Cancer Sites
And By None
And By None
And By None
And By None

Note:
To include Puerto Rico data you must select the "States and Puerto Rico" button in section 2. Selecting the "States", "Regions", or "MSA" button will exclude Puerto Rico data.

Measures (**Default measures** always checked and included. Check box to include any others.)

- **Count**
- **Age Adjusted Rates** 95% Confidence Interval Standard Error
- **Crude Rates** 95% Confidence Interval Standard Error

Additional measure "Population" (denominator) is automatically provided in Results when rates are requested.

Title

2. Select location:

Send | Help

Click a button to select locations by State, Region, or MSA.

States | Regions | MSA | States and Puerto Rico

States

- The United States
- Alabama
- Alaska
- Arizona
- Arkansas
- California
- Colorado
- Connecticut
- Delaware

What time scale of data are included (e.g., 2010-2017)?

What documentation is available for these data?

What are some of the variables included in these data?

United States Cancer Statistics Public Information Data: Incidence United States 1999 - 2018 and Puerto Rico 2005 - 2018

Summary

Summary: Cancer incidence data are available for the United States, state and metropolitan areas (MSA) by age group, race, sex, ethnicity, year of diagnosis, childhood cancer classifications and cancer site for the years 1999 - 2018. Cancer incidence data are available for Puerto Rico by age group, sex, year of diagnosis, childhood cancer classifications and cancer site for the years 2005 - 2018.

Source: The United States Cancer Statistics (USCS) are the official federal statistics on cancer incidence from registries having high-quality data and cancer mortality statistics for 50 states and the District of Columbia. USCS are produced by the Centers for Disease Control and Prevention (CDC) and the National Cancer Institute (NCI). For a list of all USCS contributors and partners, visit USCS Contributors.

Data are provided by:

- The Centers for Disease Control and Prevention National Program of Cancer Registries (NPCR)
- The National Cancer Institute Surveillance, Epidemiology and End Results (SEER) program

Data for years 1999-2018 are provided as reported to NPCR and SEER in the 2020 data submission.

In WONDER: You can produce tables, maps, charts, and data extracts. Obtain incidence counts, crude rates, age-adjusted rates, with 95% confidence intervals and standard errors for rates. Select specific disease and demographic criteria to produce cross-tabulated incidence measures. Data are organized into three levels of geographic detail: national, state and Metropolitan Statistical Areas (MSAs). The population estimates used as the denominator for rate calculations are also shown. You can limit and index your data by any and all of the variables:

1. Location - Regions, Divisions and States or Metropolitan Areas (MSA)
2. Year - 1999-2018 (2005 - 2018 for Puerto Rico)
3. Age Group
4. Race - All, American Indian or Alaska Native, Asian or Pacific Islander, Black or African American, White, Other Races Combined (not available for Puerto Rico)
5. Sex - Female, Male
6. Ethnicity - Hispanic, Non-Hispanic, Unknown (not available for Puerto Rico)
7. Leading Cancer Sites
8. Cancer Sites
9. Childhood Cancers

The following statistical measures are available as query results:

United States and Puerto Rico Cancer Statistics, 1999-2018 Incidence Request

Request Form | Results | Map | Chart | About

Cancer Statistics Data | Dataset Documentation | Other Data Access | Data Use Restrictions | How to Use WONDER

Save | Reset

*Make all desired selections and then click any **Send** button one time to send your request.*

1. Organize table layout:

Send | Help

Group Results By Cancer Sites
And By None
And By None
And By None
And By None

Note:
To include Puerto Rico data you must select the "States and Puerto Rico" button in section 2. Selecting the "States", "Regions", or "MSA" button will exclude Puerto Rico data.

Measures (**Default measures** always checked and included. Check box to include any others.)

- **Count**
- **Age Adjusted Rates** 95% Confidence Interval Standard Error
- **Crude Rates** 95% Confidence Interval Standard Error

Additional measure "Population" (denominator) is automatically provided in Results when rates are requested.

Title

2. Select location:

Send | Help

Click a button to select locations by State, Region, or MSA.

States | Regions | MSA | States and Puerto Rico

States

The United States
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware

What access restrictions exist?

What are some of the data use restrictions placed on these data?

Data Use Restrictions

Read Carefully Before Using

- **All of CDC WONDER's datasets are covered by the following policy:**
  These data are provided for the purpose of statistical reporting and analysis only. The *CDC/ATSDR Policy on Releasing and Sharing Data* prohibits linking these data with other data sets or information for the purpose of identifying an individual. If the identity of a individual described in a data set is discovered inadvertently, make no disclosure or other use of this information and report the discovery to:

  Associate Director for Science
  Office of Science Policy and Technology Transfer, CDC
  Mail Stop D50
  Phone: 404-639-7240

- **Data obtained from the National Center for Health Statistics: Compressed Mortality, Multiple Cause of Death, Linked Birth / Infant Death records and Natality, are also covered by the following policy:**
  The Public Health Service Act (42 U.S.C. 242m(d)) provides that the data collected by the National Center for Health Statistics (NCHS) may be used only for the purpose for which they were obtained; any effort to determine the identity of any reported cases, or to use the information for any purpose other than for statistical reporting and analysis, is against the law. Therefore users will:

  - Use these data for statistical reporting and analysis only.
  - For sub-national geography, do not present or publish death or birth counts of 9 or fewer or rates based on counts of nine or fewer (in figures, graphs, maps, table, etc.).
  - Make no attempt to learn the identity of any person or establishment included in these data.
  - Make no disclosure or other use of the identity of any person or establishment discovered inadvertently and advise the Director, NCHS of any such discovery.

  Confidentiality Officer
  National Center for Health Statistics
  3311 Toledo Road
  Hyattsville, MD 20782
  Phone: 888-642-4159
  Email: nchsconfidentiality@cdc.gov

- **If you belive any information obtained from CDC WONDER is incorrect, please** contact us**.**

United States and Puerto Rico Cancer Statistics, 1999-2018 Incidence Request

Request Form | Results | Map | Chart | About

Cancer Statistics Data | Dataset Documentation | Other Data Access | Data Use Restrictions | How to Use WONDER | Save | Reset

*Make all desired selections and then click any **Send** button one time to send your request.*

1. Organize table layout:

Send | Help

Group Results By: Cancer Sites
And By: None
And By: None
And By: None
And By: None

Note:
To include Puerto Rico data you must select the "States and Puerto Rico" button in section 2. Selecting the "States", "Regions", or "MSA" button will exclude Puerto Rico data.

Measures (**Default measures** always checked and included. Check box to include any others.)

- Count
- Age Adjusted Rates — 95% Confidence Interval — Standard Error
- Crude Rates — 95% Confidence Interval — Standard Error

Additional measure "Population" (denominator) is automatically provided in Results when rates are requested.

Title

2. Select location:

Send | Help

Click a button to select locations by State, Region, or MSA.

States | Regions | MSA | States and Puerto Rico

States

- The United States
- Alabama
- Alaska
- Arizona
- Arkansas
- California
- Colorado
- Connecticut
- Delaware

How else can we access the data?

U.S. Cancer Statistics Public Use Databases

Researchers can access and analyze high-quality population-based cancer incidence data on the *entire* United States population.

NPCR and SEER–U.S. Cancer Statistics Public Use Database fact sheet [PDF-170KB]

De-identified cancer incidence data reported to CDC's National Program of Cancer Registries (NPCR) and the National Cancer Institute's (NCI's) Surveillance, Epidemiology, and End Results (SEER) Program are available to researchers for free in public use databases that can be analyzed using software developed by NCI's SEER Program.

Cancer surveillance data from CDC and NCI are combined to become U.S. Cancer Statistics, the official source for federal cancer data. U.S. Cancer Statistics public use databases include cancer incidence and population data for all 50 states, the District of Columbia, and Puerto Rico, providing information on more than 33 million cancer cases.

United States and Puerto Rico Cancer Statistics, 1999-2018 Incidence Results

Request Form | Results | Map | Chart | About

Cancer Statistics Data | Dataset Documentation | Other Data Access | Help for Results | Printing Tips | Help with Exports

Save | Export | Reset

Quick Options | More Options

Top | Notes | Citation | Query Criteria

In what formats can data be exported?

Cancer Sites	Count
All Invasive Cancer Sites Combined *	30,787,702
Oral Cavity and Pharynx	759,274
Lip	42,760
Tongue	221,084
Salivary Gland	81,460
Floor of Mouth	41,574
Gum and Other Mouth	104,908
Nasopharynx	35,465
Tonsil	130,436
Oropharynx	35,159
Hypopharynx	46,339
Other Oral Cavity and Pharynx	20,089
Digestive System	5,522,754
Esophagus	319,342
Stomach	444,528
Small Intestine	144,696
Colon and Rectum	2,929,384
Colon excluding Rectum	2,110,717
Cecum	447,609
Appendix	61,983
Ascending Colon	395,400
Hepatic Flexure	101,031
Transverse Colon	190,274

United States and Puerto Rico Cancer Statistics, 1999-2018 Incidence - Notepad

File Edit Format View Help

```
"Notes" "Cancer Sites"  "Cancer Sites Code"     Count
        "All Invasive Cancer Sites Combined"    "0"     30787702
        "Oral Cavity and Pharynx"       "20010-20100"   759274
        "Lip"   "20010" 42760
        "Tongue"        "20020" 221084
        "Salivary Gland"        "20030" 81460
        "Floor of Mouth"        "20040" 41574
        "Gum and Other Mouth"   "20050" 104908
        "Nasopharynx"   "20060" 35465
        "Tonsil"        "20070" 130436
        "Oropharynx"    "20080" 35159
        "Hypopharynx"   "20090" 46339
        "Other Oral Cavity and Pharynx" "20100" 20089
        "Digestive System"      "21010-21130"   5522754
        "Esophagus"     "21010" 319342
        "Stomach"       "21020" 444528
        "Small Intestine"       "21030" 144696
        "Colon and Rectum"      "21041-21052"   2929384
        "Colon excluding Rectum"        "21041-21049"   2110717
        "Cecum" "21041" 447609
        "Appendix"      "21042" 61983
        "Ascending Colon"       "21043" 395400
        "Hepatic Flexure"       "21044" 101031
        "Transverse Colon"      "21045" 190274
        "Splenic Flexure"       "21046" 64473
        "Descending Colon"      "21047" 121169
        "Sigmoid Colon" "21048" 555566
        "Large Intestine, NOS"  "21049" 173212
        "Rectum and Rectosigmoid Junction"      "21051-21052"   818667
        "Rectosigmoid Junction" "21051" 214232
        "Rectum"        "21052" 604435
        "Anus, Anal Canal and Anorectum"        "21060" 115069
        "Liver and Intrahepatic Bile Duct"      "21071-21072"   484709
        "Liver" "21071" 424254
        "Intrahepatic Bile Duct"        "21072" 60455
        "Gallbladder"   "21080" 73949
        "Other Biliary" "21090" 109878
        "Pancreas"      "21100" 807498
        "Retroperitoneum"       "21110" 24913
        "Peritoneum, Omentum and Mesentery"     "21120" 36617
        "Other Digestive Organs"        "21130" 32171
        "Respiratory System"    "22010-22060"   4546618
        "Nose, Nasal Cavity and Middle Ear"     "22010" 44609
        "Larynx"        "22020" 250385
        "Lung and Bronchus"     "22030" 4237572
        "Pleura"        "22050" 2042
        "Trachea, Mediastinum and Other Respiratory Organs"     "22060" 12010
        "Bones and Joints"      "23000" 60240
        "Soft Tissue including Heart"   "24000" 206180
        "Skin excluding Basal and Squamous"     "25010-25020"   1401469
        "Melanoma of the Skin"  "25010" 1296399
        "Other Non-Epithelial Skin"     "25020" 105070
        "Male and Female Breast"        "26000" 4436580
```

Ln 1, Col 1 | 100% | Windows (CRLF) | UTF-8

Exercise

Work in groups to complete this exercise:

https://tinyurl.com/msswm57r

Finding More Data

Jean & Alexander Heard **Libraries**

What am I searching?

Search

DATABASES A-Z JOURNALS NEW TITLES ADVANCED SEARCH

Live Chat with the Libraries

NAME

EMAIL

STATUS

VANDERBILT SCHOOL AFFILIATION

What to Know Before You Visit

Workshops & Training

Accessibility Services

Frequently Asked Questions

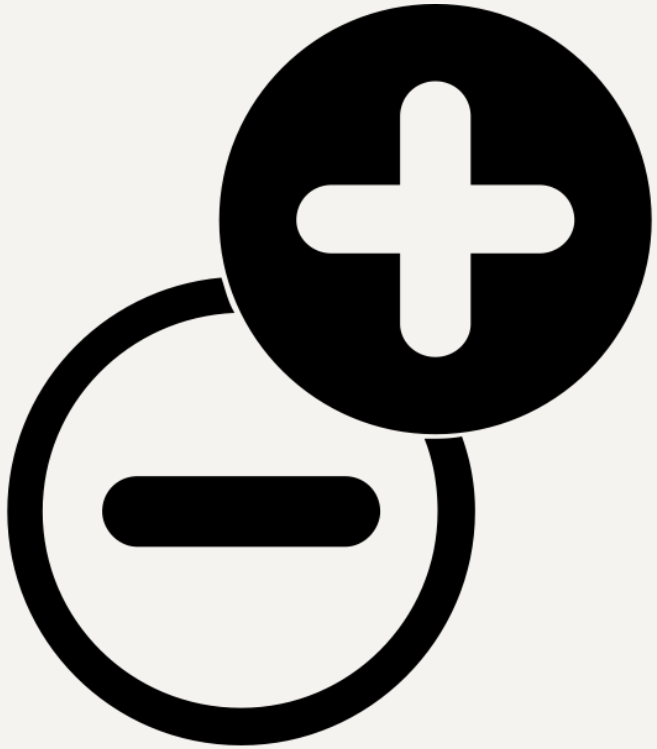
TODAY'S BUILDING HOURS

BIOMEDICAL
7:30am - 10:00pm

CENTRAL
7:30am - 12:00am

Wrap Up

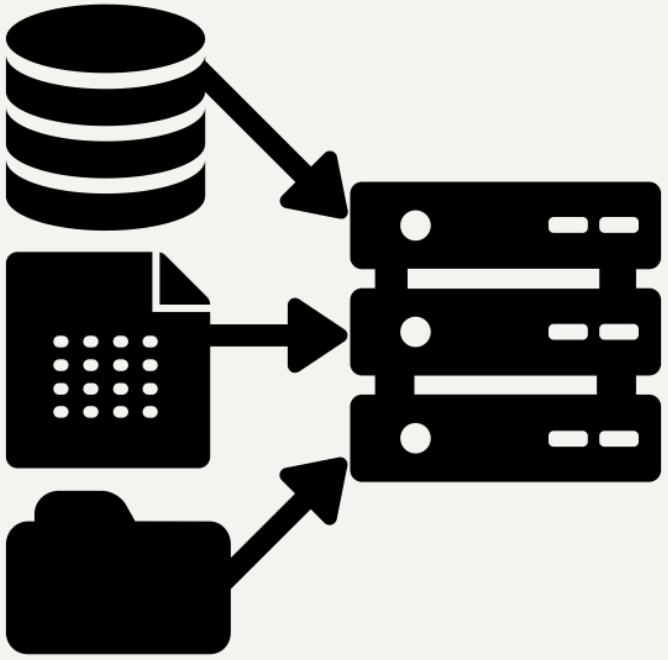
There are no perfect repositories



Created by James Kopina
from Noun Project

- Each will be better and worse in some attributes:
  - authority, accessibility, data documentation, data quality, currency, usability, interoperability
- Decide which of these is most critical for your project and your needs

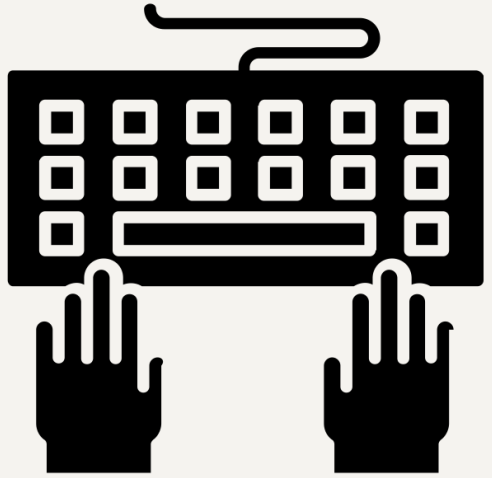
There are no perfect datasets



Created by H Alberto Gongora
from Noun Project

- The exact data you are looking for probably is not publicly available
- Researchers keep data locked down for several reasons
- “Data available upon request”
- To get the data you need, you may need to look in several places to create a mosaic

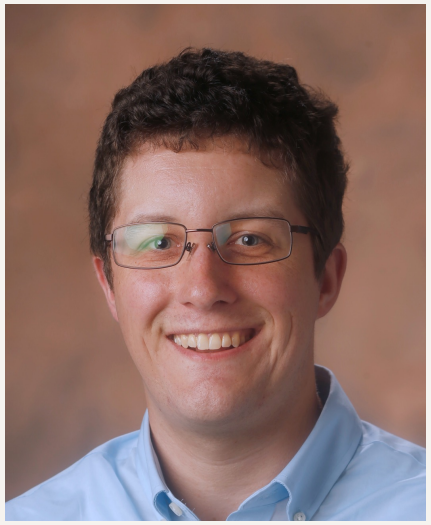
Our most important role at Vanderbilt is to help you.



Created by Vectors Market
from Noun Project

Created by Vectors Market
from Noun Project

STEM Librarian Team

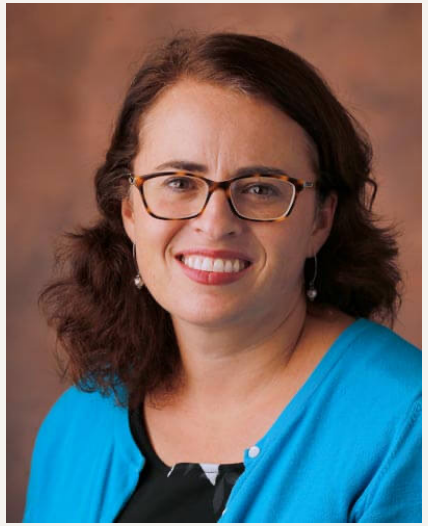


Josh Borycz
joshua.d.borycz@vanderbilt.edu

vanderbi.lt/borycz

Alex Carroll
alex.carroll@vanderbilt.edu

vanderbi.lt/carroll



Honora Eskridge
honora.eskridge@vanderbilt.edu

vanderbi.lt/eskridge

Francisco Juarez
francisco.juarez@vanderbilt.edu

vanderbi.lt/juarez

Links

BME 7410 Resource Guide: https://researchguides.library.vanderbilt.edu/bme7410_f22

These slides: https://osf.io/xpg5m