Open Data for Economic Recovery

A GUIDEBOOK BY THE BEECK CENTER FOR SOCIAL IMPACT + INNOVATION

PUBLISHED JUNE 2021

By Katya Abazajian, Natalie Ward, Kell Crowley, and Insha Momin

beeckcenter
social impact + innovation

ABOUT THE BEECK CENTER FOR SOCIAL IMPACT + INNOVATION

The Beeck Center is an experiential hub at Georgetown University that trains students and incubates scalable, leading edge ideas for social change. We believe impact at scale requires the courage to think and behave differently. Our work centers on investing in outcomes for individuals and society. We equip future global leaders with the mindset to promote outcome-driven solutions, using the tools of design, data, technology, and innovation. We convene actors across the public, private, and civic sectors to advance new tools, frameworks, and approaches necessary to achieve these outcomes.

ABOUT THE STATE CHIEF DATA OFFICERS NETWORK

The State Chief Data Officers (CDO) Network is made up of approximately 30 Chief Data Officers working in state governments. The State CDO Network works to surface and scale best practices and opportunities for collaboration across states. We also aim to support states in the creation of a CDO role. To help states improve their use of data, the State CDO Network produces core frameworks, learning opportunities, and key research to guide CDOs through structuring effective data programs.

ABOUT THIS GUIDEBOOK

This guidebook was compiled by staff of the State CDO Network at the Beeck Center for Social Impact + Innovation. We compiled these findings based on interviews with subject-matter experts and conversations with members of the State CDO Network. We used this research to identify data sets that states can and should publish to increase transparency and accountability and effectively respond to the COVID-19 crisis. The key datasets in this report are coordinated with issues and use cases developed for the State CDO Network's previous report, Leveraging Data for Economic Recovery: A Roadmap for States. Additionally, we conducted research on the legal and policy environment around open data in states, and the range of state-level decisions on the privacy of data related to COVID-19. This report is published with support from the Bill & Melinda Gates Foundation.

Table of Contents

Introduction

Open data initiatives are a fundamental component of good government that can benefit public servants and residents alike. Open data helps to bust silos within government and creates opportunities for community members and civil society to participate in public policy. While they may seem like just one small part of a state government's data strategy, open data initiatives are the key to unlocking information that can help answer challenging policy questions.

Anyone can use open data. Public servants at the state level might use it to discover what data other agencies have and are able to share. Policymakers might use it to understand and inform upcoming policy decisions. Advocacy organizations might access it to improve the quality of their research and representation of communities' needs. And residents might access it to look for specific information about issues facing their communities. Publishing the right open data requires first understanding the purpose.

Through establishing robust open-data portals, chief data officers (CDOs) have made strides toward modernizing, streamlining, and creating a culture of transparency within their state governments. However, the roles of CDOs, including their influence over funding or decision-making about open data initiatives, can vary. Because capacities vary state by state, it is important for CDOs to be aware of the best-of-breed datasets their colleagues are publishing. Knowing which datasets are critical to publish on open data portals makes advocating for their publication across state governments easier. Open data can further transparency and evidence-based policy making within any state government, therefore, it is imperative that CDOs begin to cultivate streamlined, robust, and thorough open data portals.

To achieve better outcomes for residents during the phases of pandemic recovery, states must begin to leverage data and post open data that goes beyond baseline COVID-19 summary statistics for the state. As recovery progresses, it is vital that states publish data not only directly related to COVID-19—such as more granular testing and case data—but also datasets that demonstrate the economic and social effects of COVID-19, including applications for SNAP and TANF, unemployment, business openings and closures, or foreclosures and evictions.

Getting Started

WHAT IS OPEN DATA?

Open data is any public data or information that is publicly available without fees or restrictions on use. At its best, open data is accessible online, machine-readable, downloadable in bulk, well-documented, and usable in a variety of formats.

In 2017, the Sunlight Foundation published the 10 most popular open dataset topics in cities, which included information on police and crime, transportation, finance, and elections, among other topics. States can learn from how open data has evolved in cities—from publishing fundamental open datasets that create transparency and access to information, to supporting communities in using that data for social good.

Research shows that when open data policies and practices are robust, and when states publish the data that users actually want, governments save time and money on responding to public record requests. An **open data policy** is any legislation or administrative order that mandates the proactive disclosure of non-sensitive information. States can maximize the benefits of open data by thoughtfully prioritizing which datasets to publish and passing comprehensive open data policies.

The Sunlight Foundation has also published open data policy guidelines for governments to implement effective open data legislation. These guidelines, originally written for cities, help lawmakers decide which data should be made public, how to make that data public, and how to implement open data policy.

At a minimum, open data policies can help governments establish open data programs and mandate the proactive disclosure of non-sensitive information. They can require strong metadata and comprehensive formats, extend open data requirements to third party contractors, and establish open data oversight authorities and timelines. These policies can also build on existing public accountability policies (like campaign finance regulation).

The National Conference of State Legislatures (NSCL) published an outline of the common features across the current ecosystem of state-level open data laws. These are similar to city-level best practices: designate a CDO, require the use of machine-readable formats, establish open data programs, and create oversight authorities and governance structures. By following these guidelines, states can launch and maintain effective and impactful open data programs.

When launching open data initiatives, CDOs can build on the lessons learned from other cities and states implementing best practices. Lessons can help states tackle the basics of publishing usable and effective open data, or help them improve open data to be more targeted and impactful in residents' lives.

THE OPEN DATA BASICS

Publish APIs.	Application Programming Interfaces (APIs) are standard features in most off-the-shelf open data portals, and are worth the effort even in custom solutions because they allow researchers and other data users to bulk download information on an on-going basis. Like any tools, APIs need good documentation to maximize the reuse of public data.
Coordinate with public records offices.	The most obvious economic value-add of open data on the government side is the reduction of time and energy spent responding to public records requests once robust open data portals are launched. Directly linking public records processes to the prioritization process for publishing data can help open data programs demonstrate early results..
Convert PDFs.	Publishing PDFs is a safe practice for government agencies that are reporting data about operations that have previously taken place through paper-based workflows. Many feel that PDFs protect the provenance of their public data. But searchability and machine-readability are key features that make data usable by people who are looking for specific pieces of information or looking to download and analyze data in bulk. Converting PDFs to machine-readable formats can ensure that open data programs deliver on innovation.
Create robust documentation.	Documenting datasets and producing data libraries can be essential for encouraging responsible reuse of public data both inside and outside of government. Documentation protocols can include assessments of privacy risk or potential biases in the data, allowing public data users to understand the appropriate uses of data and minimizing risk or harm. Most data owners in government are under-resourced for robust data documentation and stewardship but these are essential foundations of open data programs.
Make data discoverable.	Colf data can't be accessible, it should at least be discoverable. Making data discoverable means getting agencies to complete data inventories and publishing them so that members of the public can see what information lives where. States like Connecticut and Virginia publish data inventories that are updated and improved over time to ensure that people can see what data the state is holding and even what protections it might be subject to. Inventories are an important first step to setting up a robust data governance strategy and open data program.

IMPROVING OPEN DATA

Disaggregate data for analysis.	Users often need granular demographic data in order to understand the effects of public programs on specific populations. Data that is disaggregated by race or ethnicity can be aggregated or anonymized at appropriate geographic levels in order to provide insights while still protecting personal privacy. Disaggregating open data by race and ethnicity is essential for understanding disparate impacts of public policies and services.
Follow a beat.	Publishing 500 datasets at once can be a quick way to put stress on agencies and overwhelm public data users. Data users inside and outside of government benefit from curation. Publishing one dataset a month or datasets related to one topic per month can be an effective strategy for letting the public know what's on the horizon and which datasets are coming down the pipeline.
Open doors for researchers.	Academic or industry researchers are common consumers of state-level data. Often, researchers will go through backchannels to attempt to establish data-sharing agreements or will submit public records requests for valuable information. Improving relationships with researchers and designing special pathways to collaboration like "data cleanrooms," which only share a subsection of data with outside users, can build trust and ensure that state public data is useful.
Create robust documentation.	Documenting datasets and producing data libraries can be essential for encouraging responsible reuse of public data both inside and outside of government. Documentation protocols can include assessments of privacy risk or potential biases in the data, allowing public data users to understand the appropriate uses of data and minimizing risk or harm. Most data owners in government are under-resourced for robust data documentation and stewardship but these are essential foundations of open data programs.
Know your audiences.	Conducting interviews or focus groups with key users can help data owners at state agencies to understand who might be using their data and how they might improve its usability or accessibility for that audience. Users can be from other government agencies or at sub-state levels. In any case, engaging users directly, even in small groups or through ad hoc outreach, can help data owners make better planning decisions about open data and proactively identify chances for impact.

IMPLEMENTING DATA PROTECTIONS

Open data program administrators are tasked with gathering all of the data that is fit to be published, and ensuring that none of the data poses a risk to individuals represented in that data. Often, any sensitive data in a raw dataset is removed before publication. But after data is deemed non-sensitive and published, the main risk to publishing open data is exposing individuals to **re-identification risk** in sensitive situations, or publishing data that users can manipulate to reveal personally identifiable information.

Because many open data programs are still in their first few years of existence, there haven't been significantly harmful cases of re-identification of individuals through open data. Naturally, there are exceptions. Sometimes, governments can suffer data leaks or accidentally publish information that was intended to be kept secret, exposing individuals to risk. In general, states need to prepare for the reality that as more data becomes publicly available, individuals will be more at risk of re-identification through public sources.

The most common strategy for protecting potentially identifiable data when publishing open data is **aggregation**. Open data administrators should use standardized protocols and quality checks to determine at which geographic level data points should be aggregated. For example, if publishing data on food assistance that is reported at the Census tract level, are there any Census tracts where only one or two individuals are the only recipients of assistance? Could a data user identify those people by cross-referencing food assistance data with Census tract-level locations of rental assistance? Connecticut's open data program publishes guidance to help open data administrators make these calls.

Other strategies for protecting sensitive data in geographic formats could include randomizing the locations of points on a map within a certain geographic boundary. For example, placing points that identify individuals at the center of a Census tract rather than directly on the location of the individual's address. Alternatively, publishers could simply redact the specific address or coordinates, and provide a geographic identifier such as a zip code, locality, or Census tract.

In states where state agencies are in charge of their own open data portals or programs, it can be more difficult to cross-reference datasets and assess re-identification risk. Centralized open data programs or guidance for data protections can help to avoid privacy risks across agencies.

Privacy Environment and Key Legislation

Most open data laws include protections for sensitive data that allow officials to weigh whether the public risk of publishing certain data would outweigh the potential public benefit. Attorneys general and public records officials are often those who decide whether certain datasets would present a potential harm if published. However, a general misreading of rules of thumb about sensitive data means that governments can sometimes protect data that would harm the government if published, rather than weighing what might harm members of the general public.

Balancing openness and privacy is an inherent challenge that requires careful planning and data governance structures that help people running open data programs decide what can be shared proactively. The foundation of personal privacy laws in the U.S. is the Privacy Act of 1974, which created the modern standards for Personally Identifiable Information (PII) and Personal Health Information (PHI).

The OPEN Data Act, passed in 2017, requires agencies to make government data available in machine-readable and open formats by default under open licenses. According to the Act, "nonpublic data assets":

> (A) means a data asset that may not be made available to the public for privacy, security, confidentiality, regulation, or other reasons as determined by law; and
>
> (B) includes data provided by contractors that is protected by contract, license, patent, trademark, copyright, confidentiality, regulation, or other restriction.

This definition leaves significant room for interpretation across federal agencies and state and local governments following suit. As a result, states have made differing decisions about the sensitivity of public data to be released, especially under the umbrella of data related to COVID-19. States have taken varying measures to address the specific privacy risks related to contact tracing, which generally hold little bearing on the privacy or availability of public data. As COVID-19 has put a spotlight on governments' roles in collecting and protecting data, state public officials should be aware of the existing constraints to data-sharing and the opportunities for stronger data protections across state agencies.

Rather than focusing on new protections, some states, like Indiana, have responded to COVID-19's new questions around privacy and access to information by creating new ways to collaborate across sectors. Indiana's state analytics office, called the Management Performance Hub, created a "sandbox" for researchers to access the state's COVID-19 data in a safe yet accessible manner. The Enhanced Research Environment allows the state to bypass lengthy MOU processes and focus on sharing data with trusted individuals who could support the state's public health efforts.

KEY LEGISLATION

The unprecedented interconnectedness of data and efforts to streamline and centralize public services raise new questions about the privacy of individuals who engage with federal, state, and local governments. The strongest legal data protections in the U.S. are the Health Insurance Portability and Accountability Act (HIPAA) (1996) and the Family Educational Rights and Privacy Act (FERPA) (1974), which were passed at the federal level and affect all health providers and educational organizations. In recent years, states have begun passing their own privacy laws to address emerging technologies and mass data collection in the consumer space like the California Consumer Privacy Act and the Illinois Data and Transparency Privacy Act.

HIPAA

The U.S. Department of Health and Human Services (HHS) publishes specific guidance to address de-identification, or removal of personal information, of HIPAA-protected data. HIPAA specifically protects personal health information (not any personally identifiable information), so open data administrators can detach personally identifiable information from health information and still be compliant with HIPAA. The guidance is as follows: "Protected health information exists when we can link an individual to their personal health information. Identifying information alone, such as personal names, residential addresses, or phone numbers, is not designated as PHI. Nor is a report that only contains the average age of health plan members;"

In addition to allowing the de-identification of HIPAA data, this guidance also clarifies that the HIPAA Privacy Rule no longer applies after the fact: "Once de-identification is achieved, the Privacy Rule does not restrict the use or disclosure of the health information, as it is no longer considered protected. There are two ways to de-identify protected data. The first is to have a qualified expert review the requested information and go through a formal statistical process to determine the risks associated with releasing the information." At the state level, Chief Data Officers are considered qualified experts who can confirm through statistical analysis whether there is any risk in releasing de-identified information (which they can do using analysis based on the previous section in this report on re-identification risk).

Updates related to COVID-19: HHS publishes the latest information about HIPAA guidance on an on-going basis. At the start of the pandemic, HHS expanded the ability of HIPAA-covered entities to meet the needs of medical patients, for example by allowing wider use of telehealth technologies. In February, HHS and its Office of Civil Rights published a joint bulletin explaining how HIPAA covered entities and their business associates should handle notification and publication of COVID-19 cases.

Importantly, the bulletin includes a recommendation to disclose the "minimum necessary" amount of information and to look to public health authorities to determine how much information is necessary. From the bulletin: "For example, a covered entity may rely on representations from the CDC that the protected health information requested by the CDC about all patients exposed to or suspected or confirmed to have Novel Coronavirus (2019-nCoV) is the minimum necessary for the public health purpose."

FERPA

FERPA covers schools and the protection of education records that might connect students to information about their families and other personal records. While aggregate test scores and other general summary statistics are normally allowed under FERPA, privacy rules exist to protect students, children, and their families. HHS provides guidance on the overlap of FERPA and HIPAA data in the event that agencies handle student health data. Importantly, HHS has published guidance in the past noting that primary and secondary schools are not HIPAA-covered entities unless they specifically employ an operating healthcare provider, meaning that, by-and-large, primary and secondary schools may defer to FERPA privacy rules.

Updates related to COVID-19: Agency guidance from March states that FERPA does not block schools from publishing data on COVID-19 cases, only if data is about a specific individual and only if the information is being shared openly (as opposed to with limited discretion). "As an example, the memo said schools might need to share identifiable information, including the affected person's name, with members of a wrestling team if someone on the squad contracted COVID-19."

BALANCING PRIVACY AND TRANSPARENCY IN COVID-19

The COVID-19 pandemic placed unprecedented scrutiny on state data practices. National media have urged states to publish "complete and accurate information" about the pandemic. Requests about COVID-19 testing, immunizations, and other issues of public health are not likely to stop in the coming years as the nation navigate recovery.

But beyond their general efforts to keep data available, accurate, and up-to-date, states have taken differing approaches to publishing granular COVID-19 testing data, school re-opening data, or data tracking community spread. For example, New York local governments began publishing case counts at a county level, but after public pressure released statistics at a Zip code level instead. Calls for more granular data on COVID-19 and demographic impacts were mirrored around the country with most governments eventually responding to public pressure and publishing testing data on a neighborhood or zip code level. But other issues also arose around schools selectively using FERPA to avoid publishing statistics, states hiding behind Freedom of Information Act (FOIA) exemptions for epidemiological investigations, governments struggling to accurately report on nursing homes, and struggles to hold private companies accountable when they are engaged in public health coordination.

Is COVID-19 testing data protected under FERPA?

An August 2020 report by *USAToday* noted that schools in Tennessee, Florida, Indiana, and Missouri had been citing HIPAA and FERPA to rationalize obscuring information about COVID-19 cases on campus. *The Washington Post* also reported that public universities—including the University of Alabama, Arizona State University, the University of Houston, and the University of North Carolina at Chapel Hill—cited both privacy laws to avoid discussing student health. Schools and universities are not HIPAA-protected entities, and therefore do not have to comply with HIPAA protections around student health. The Department of Education released special guidance in March 2020 stating that in the current crisis, schools have a responsibility to publish case counts and other COVID-19 data, and can do so without publishing students' personal records. Under the Clery Act, schools are also required to report campus safety threats, which captures crises like the pandemic.

Should states use FOIA exemptions on public health data?

States like Georgia and Oregon have used specific FOIA exemptions to withhold public information about COVID-19. While both states eventually published COVID-related information to meet public demands, the FOIA exemption exposed a loophole in public information law. Georgia initially cited a FOIA exemption about "private commercial information" to withhold information about where protective equipment was being shipped in the state. But, according to research by the Open Contracting Partnership, "commercial information cannot be legitimately sensitive if it's already known to competitors in some jurisdictions," and "even commercially sensitive information may be disclosed based on a public interest test." Oregon cited an exemption intended for ongoing epidemiological investigations, which says that the state can deny public information requests based on the discretion of public health officials, but can still deem it necessary to release in the public interest. Following public pressure, the state began publishing weekly COVID-19 metrics at more granular levels.

Are people in nursing homes at risk of re-identification?

In May 2020, the HHS published COVID-19 reporting guidance for nursing home facilities. Facilities must electronically report information about COVID-19 in standardized formats, including staffing shortages, ventilator capacity, suspected and confirmed cases, and deaths. Yet as of May 2020, 13 states were not publishing data on nursing homes or long-term care facilities. San Diego County initially refused to publish information about assisted living facilities with fewer than six beds citing privacy concerns. Arizona officials said releasing data on nursing homes "could create stigma" and other states like Mississippi and Virginia leaned on similarly general but not legally founded privacy concerns. States like Pennsylvania eventually did publish data on nursing homes after public pushback.

Which rules for public health data apply to private companies?

Private companies must be held accountable to HIPAA standards if cooperating with public health agencies as business associates. Business associates can risk an investigation by the HHS's Office of Civil Rights if they are responsible for deploying apps that fail to protect PHI. If, for example, a company considered is a health agency business associate in charge of developing a contact tracing app, then HIPAA's Privacy Rule and Security Rule may apply. Companies must follow recommended processes to aggregate, anonymize, and/or de-identify any data stored.

How should governments navigate specific privacy questions?

States may not always have the answers right away to determine the privacy risks of publishing open data. Chief Data Officers can serve as qualified experts in helping to de-identify, aggregate, and otherwise protect health-related data to be shared or published. State decision-makers can also continue looking to research about disclosing data related to specifically sensitive policy issues like domestic violence or opioid/substance abuse. For example, the Network for Public Health Law published guidance for domestic violence agencies or shelters looking to publish COVID-19 reporting data. Agencies may report cases of COVID-19 but may not disclose personally identifying information about a service recipient in connection with that reporting unless the agency has obtained consent or the disclosure is mandated by a statute or court. Listening to affected communities and engaging experts to determine evolving requirements around sensitive policy areas can help state decision-makers stay up-to-date on privacy needs.

Critical Cases in States

The State CDO Network has 30 member states with established CDOs, all of which publish some kind of open data. Of the state open data programs in the State CDO Network, 22 are operated by the IT or executive agencies where CDOs sit. Seven others are operated by specific state agencies that run their own open data portals. About 17 of the 30 member states publish critical data that goes beyond basic geospatial data. Critical data is data that goes beyond establishing operational transparency and accountability in states, additionally supporting critical efforts to implement policy or affect reforms.

View our gallery of State Open Data Portals

PUBLIC ASSISTANCE CASES IN CONNECTICUT

The Connecticut Department of Social Services (DSS) publishes weekly and daily DSS Application Activity Before and During COVID-19 Emergency. The dataset includes the paper and online applications submitted directly to the Department of Social Services by day. Data includes:

- Total applications received (includes multi-program apps)
- Cash assistance applications (TFA, SAGA, State Supp)
- Medical applications (HUSKY C, LTSS, MSP)
- SNAP applications

UNEMPLOYMENT INSURANCE CLAIMS DATA IN CALIFORNIA

COVID-19 shut down the economy for more than a year and left millions of Americans out of work. An historic number of Americans sought out unemployment benefits after the onset of the pandemic, and will likely continue to struggle as the economy slowly returns to pre-pandemic levels. California publishes weekly unemployment insurance claims data which helps to monitor unemployment trends in the state. Initial claims measure emerging unemployment and continued weeks claimed measure the number of persons claiming unemployment benefits.

RENT RELIEF PROGRAM DATA IN MARYLAND

Across the country, states are addressing the most critical housing needs in their communities. Renters, landlords, and homeowners are struggling to make housing payments during this time. In Maryland, the state COVID-19 dashboard has been a pivotal element of the Rent Relief Program. The U.S. Department of Health and Human Services has developed a Homeless Prevention Index to evaluate all neighborhoods in relation to COVID-19 impact, housing stress, and social determinants. Applications for rent relief for Maryland citizens are prioritized based on areas of highest impact, with households outside of the initial "high impact" neighborhoods placed on a waitlist and contacted as capacity allows.]

The Top 20 Open Datasets in States

These lists, presented in alphabetical order, is a point-in-time snapshot of open datasets that create a strong foundation for state open data portals or are directly relevant to economic recovery from COVID-19. This list was compiled based on a scan of datasets available on existing state open data portals and interviews with open data experts working at the state level.

TIER 1: FUNDAMENTAL DATASETS

Dataset	Description	Timeliness
Administrative Boundaries	District maps for schools, voting, etc.	Annual
Broadband Infrastructure	Broadband accessibility and speed data.	Annual
Budget and Spending	Budget and spending amounts by department.	Monthly
Business Registrations	Licenses and registrations for businesses.	Monthly
Housing Markets	Housing sales and/or assessments.	Varies
Population Statistics	Demographics and Census data.	Annual
Procurement and Contracts	Solicitations, costs, and executed contracts.	Weekly
Property Ownership	Records of parcels and property owners.	Annual
Tax Revenues and Forecasts	Economic forecasts and revenue reporting.	Monthly
Unemployment	Unemployment insurance and claims.	Weekly

TIER 2: CRITICAL DATASETS

Dataset	Description	Timeliness
Career/Technical Education	CTE providers, programs, and outcomes.	Annual
Childcare Providers and Slots	Accredited providers with capacity and resources.	Monthly
COVID-19 Impacts	Testing, vaccination, and health impacts.	Varies
Emergency Shelter Use	Capacity at shelters and emergency shelters.	Daily
Evictions	Evictions data from courts, rental assistance.	Weekly
Food Assistance Uptake	Food programs, SNAP, and farmers markets.	Monthly
Healthcare Claims	Medicaid, Medicare, and all-payer claims.	Monthly
Opioid Crisis	Opioid-related policy and public health outcomes.	Monthly
Public Assistance Cases	All requests for benefits and public assistance.	Weekly
Public Health Outcomes	Public health statistics or program reporting.	Varies

FUNDAMENTAL DATASETS

Open data was originally established as a tool for the public to engage with governments in monitoring legislative activity, analyzing non-sensitive data, and generally ensuring that the public can hold governments accountable. Datasets that are fundamental to public transparency and accountability should be published openly regardless of specialized use because public access to information creates a general sense of trust and participation among residents. The following datasets are fundamental in that they provide an essential view into public operations and allow for residents to have ongoing access into foundational public services and operations.

Business Registrations

Data about business registrations primarily contains business licenses, incorporations, or new business registrations. Data about businesses should include robust metadata stating how often data is updated, and through which means the state collects data on businesses. In many states, this data is compiled by the Secretary of State's office.

Timeliness: Monthly

Formats: This data is often presented in a search tool but can also be published in a searchable dataset.

Risk level: Low. Data is often non-sensitive public record and includes minimal personally identifiable information.

Connection to federal programs: The Census Bureau connects relevant business registration data by compiling various data sources, including Employee Identification Numbers from the internal Revenue Service (IRS). The Census business register is protected by Title 13 and Title 26, US Code.

Examples:

Dataset	State	Sample Fields
New Businesses Registered	Oregon	Business Name, Entity Type, Registry Date, Name, Address
Corporations Data	Washington	Unified Business Identifier, Business Name, Record Status, Date of Incorporation, Dissolution Date, Business Type
Business Entities	Colorado	Business Name, Address, Entity Status, Jurisdiction, Entity Type, Agent Information

Property Ownership

Real property data consists of a record of properties, their characteristics (use class, land use, lot size, etc.), and their owners. Often, data on property ownership is published at the local or regional levels when it is not available at the state level.

Timeliness: Annual

Formats: Making data available by bulk download or API is important for users of property ownership data. Search tools can also help users find specific properties. .

Risk level: Medium-low. In many places, property ownership data is already available through proprietary data intermediaries including names and business affiliations of property owners.

Connection to federal programs: N/A

Examples:

Dataset	State	Sample Fields
Landlord Ownership/ Address Tool	Pennsylvania	Address, Use Class, Owner Type, Land Use, Lot Size, Property Owner, Assessment Values, Previous Sales, Foreclosure Filings
Real Property Search Tool	Maryland	Use, Owner, Address, Parcel, Neighborhood, Property Land Area, Value, Seller History

Budget and Spending

Data about budget and spending primarily contains information about how funds are allocated for various state agencies and how those agencies spend their funds. Budget data can refer to budget allocations mapping where funds are intended to be spent for the fiscal year. Spending data can refer to audited data on expenses at the department level or more granular spending like contract payments.

Timeliness: Monthly

Analysis required: This data requires gathering information from across departments and processing it into usable or readable data about budgets and spending.

Risk level: Low. No personally identifiable information or sensitive business information.

Connection to federal programs: N/A

Examples:

Dataset	State	Sample Fields
Ebudget	California	State Agencies, Positions, General Fund, Special Funds, Bond Funds
Spending/Checkbook	Delaware	Spending, Transactions, Vendors, Expenses, Expense Category, Departments and Agencies, Payments Over Time
Active Pension Members	New Jersey	Member Name, Enrollment Date, Pension Fund, Member Retirement Cause, Total of Monthly Pension Allowance, Last Employer Information

Population Statistics

Data about population statistics includes the most recent demographic estimates and projections broken down at least by place, age, sex and race/ethnicity.

Timeliness: Annual

Formats: This data may not need to be published in the form of a dataset, given that raw data is often available directly from the Census Bureau. Visualizing or publishing data may require integrating Census data into state systems.

Risk level: Low. This data should disclose minimal personally identifiable information.

Connection to federal programs: This data is compiled by the Census Bureau.

Examples:

Dataset	State	Sample Fields
Census/Demographics	Texas	Population, Households, Race and Ethnicity, Age, Families, Sex, Children, Tenure, Group Quarters

Tax Revenues and Forecasts

Tax revenue data and economic forecasts incorporate economic industry outlooks, taking into account employment demand, consumer confidence metrics, and price inflation, as well as data about tax rates and revenues.

Timeliness: Monthly

Formats: Analysis may be required to convert IRS or tax reporting data into forecasts relevant to specific industries or regions.

Risk level: Low. This data discloses minimal personally identifiable information.

Connection to federal programs: N/A

Examples:

Dataset	State	Sample Fields
State and Local Purpose Taxes and Fees	New York	Fiscal Year, Tax Type, Month of Collections, Amount Collected
Tax Revenues	Connecticut	Month, Calendar Year, Fiscal Year, Withholding, Income Tax Estimates & Finals, Repealed Taxes, Total Revenue
Regional Economic Analysis Profiles	California	Industry, Occupations, Job Openings, Wages, Education Level, Work Activities, Job Advertisements, Employment Percentage
Key Economic Indicators	Texas	Consumer Confidence Index, Unemployment, Personal Consumption Expenditure, Consumer Price Index, Gasoline and Diesel Prices, Building Permits and Construction, Sales Tax Collections

Housing Markets

Housing markets data may consist of sales, rental information, mortgages, titles, and other housing-related assessments.

Timeliness: Varies

Format: Data should be available for bulk download or available via serach tool with a bulk download option.

Risk level: Medium-low. May include housing addresses but minimal personally identifiable information.

Connection to federal programs: Home Mortgage Disclosure Act (HMDA) data is provided by the Consumer Financial Protection Bureau (CFPB).

Examples:

Dataset	State	Sample Fields
Sales	Connecticut	Address, List Year, Assessed Value, Sale Amount, Sales Ratio, Property Type, Assessor Remarks
Affordable/Assisted Housing	Connecticut	Town, Census Units, Government Assisted, Tenant Rental Assistance, Single Family Mortgages, Deed Restricted Units, Percent Affordable
Mortgage Loans Purchased	New York	Bond Series, Original Loan Amount, Loan Purchase Date, Original Loan to Value, Loan Type, Amount, Property Type, Housing Size

Unemployment Claims

This data includes records of unemployment claims and counts of individuals receiving regular unemployment insurance benefits.

Timeliness: Weekly

Format: Summary data on claims must be aggregated out of individual claims data in order to protect individuals.

Risk level: Medium. Raw data includes sensitive, personally identifiable information and must be anonymized before publication.

Connection to federal programs: The unemployment insurance program is run by states in partnership with the U.S. Department of Labor. Additionally, the Bureau of Labor Statistics supports state efforts to create Labor Market Information sysmtes for the specific prupose of making this data available.

Examples:

Dataset	State	Sample Fields
Weekly UI Claims	California	Area, Initial Claims, Continued Claims, Covered Employment, Insured Unemployment Rate
Monthly UI Claims	New York	Year, Month, Region, County, Beneficiaries, Benefit Amounts
Monthly UI Claims	Montana	Year, Period, Area, Industry, Occupation, Claimants

Administrative Boundaries

This data includes GIS data on administrative boundaries for school districts, voting districts, police districts, and any other geographic boundary files.

Timeliness: Annual

Format: GIS files should be available for direct download, but can also be portrayed via mapping or visualization.

Risk level: Low. This data does not include personally identifiable information.

Connection to federal programs: The U.S. Census Bureau routinely collects administrative and other boundary data from states both as part of the decennial Census and ongoing efforts to define statistical areas.

Examples:

Dataset	State	Sample Fields
Schools Map	Connecticut	Name, Organization Type, Address, Open Date, Magnet Status, Grades Offered
Administrative Boundaries	New York	Boundaries of state, counties, cities, towns, villages, and Indian territories

Broadband Infrastructure
This data consists of broadband availability, adoption, and infrastructure specified by location, whether in the form of maps, dashboards, or datasets.

Timeliness: Annual

Format: Data from broadband speed surveys is often available via mapping and is aggregated to the summary level by Census tract, neighborhood, or other geography.

Risk level: Low. This data does not include personally identifiable information.

Connection to federal programs: The Federal Communications Commission (FCC) tracks Form 477 data which attempts to gather information on connectivity speeds. The Broadband DATA Act was passed in March 2020 to improve the quality of the FCC's data collection on broadband accessibility.

Examples:

Dataset	State	Sample Fields	Type
Adoption and Infrastructure Map	California	Median Household Income, Broadband Adoption, Urban/ Rural, Political Boundaries, Fixed Served Status, Infrastructure Eligible Areas, Wireline and Wireless Served Status	Map
Infrastructure by Municipality	New York	Municipality, Housing Units, Cable Providers, Digital Subscriber Line Providers, Fiber Providers, Wireline Providers, Wireless Providers, Satellite Providers	Dataset
Infrastructure and Speed Map	Georgia	Served and Unserved Areas, Percent Unserved	Map

Procurement and Contracts

This data includes registered vendors, active suppliers, purchase orders, open solicitations, closed solicitations, executed contracts, and contract payments. Data on solicitations and executed contracts should include contract amounts, timelines, and vendor information.

Timeliness: Weekly

Format: Procurement data may originally be in PDF form but should be processed into machine-readable formats connected by unique identifiers by contract number.

Risk level: Low. This data does not include any personally identifiable information. Data owners should ensure that vendor data is openly available.

Connection to federal programs: N/A

Examples:

Dataset	State	Sample Fields
Contracts and Procurements Registration	California	Department Name, Supplier Name, Item Name, Grand Total, Acquisition Type, Buyer Name, Status
Orders, Suppliers, and Contracts Search Tool	Virginia	Contract Officer, Authorized Entities, Agency, Total $, Contract Number, Date, Supplier and Buyer Address

CRITICAL DATASETS

Career and Technical Education

Data about career and technical education (CTE) outcomes includes information about providers, programs, and student outcomes either at the individual or summary level.

Timeliness: Annual

Formats: CTE data can be found in some education longitudinal data systems. Otherwise, individual-level data is compiled directly from providers and must be aggregated into tabular summary-level data.

Risk level: Medium-high. Raw data includes personally identifiable information, is protected by FERPA, and should not be published as open data. Data must be aggregated and anonymized before publication.

Connection to federal programs: The Workforce Innovation and Opportunity Act (WIOA) requires alignment between workforce development programs and local CTE infrastructure partners. This can include standardizing data collection and reporting on CTE programs.

Examples:

Dataset	State	Sample Fields
Annual CTE Outcomes	Pennsylvania	CTE Enrollment, Industry Certifications Offered, Postsecondary Transition, Employment Rate, Median Earnings, Measurable Skill Gains

COVID-19 Impacts

Data about COVID-19 impacts includes baseline information on COVID-19 cases, health risks, economic risks, and other impacts, disaggregated by age, gender, race, and ethnicity. States are publishing disaggregation of the baseline information by county, town, or zip code, including data on COVID-19 stimulus funds.

Timeliness: Varies

Formats: This data is often published by dashboard or visualization to communicate general trends and analysis across demographic groups.

Risk level: Medium-low. In some cases, disaggregating data on COVID-19 cases can indirectly identify individuals, especially in small geographies. Take appropriate steps to anonymize and mask individual health data. Other disaggregated data on COVID-19 impacts should contain minimal personally identifiable information.

Connection to federal programs: The Coronavirus Aid, Relief, and Economic Security Act and American Rescue Plan Act both delivered significant stimulus to state and local governments. Funding allocations and outcomes should be reported using data disaggregated by race, gender, and other demographic characteristics.

Examples:

Dataset	State	Sample Fields
New York Open Data Portal	New York	Statewide COVID-19 Testing; COVID-19 Outcomes by Testing Cohorts: Cases, Hospitalizations, and Deaths; COVID-19 Daily Counts of Cases, Hospitalizations, and Deaths
Connecticut Open Data Portal	Connecticut	COVID-19 Tests, Cases, and Deaths (By Town); COVID-19 Cases and Deaths by Age Group; COVID-19 in PK-12 Public and Private Schools
Virginia Open Data Portal	Virginia	COVID-19 Vaccine Phase By HealthDistrict; COVID-19 Outbreaks; COVID-19 Tests by Lab Report Date

Childcare Providers and Slots

This data consists of accredited, open, and active childcare providers with capacity and childcare resources.

Timeliness: Monthly

Formats: Data on childcare providers is published at the provider level. It should be searchable for users to find specific providers as well as being available for bulk download.

Risk level: Low. Data does not include any personally identifiable information.

Connection to federal programs: The Office of Head Start programs at the Department of Health & Human Services collects data on Head Start and Early Head Start providers. This includes data on open and accredited childcare providers.

Examples:

Dataset	State	Sample Fields
Pre-K and Head Start Slots	Pennsylvania	Address, Location Served, Full Day Funded Spots, Half Day Funded Spots
Open and Active Childcare Providers	Pennsylvania	Facility Name, Provider Type, Facility Address, STARS Level, Current Status
Registered Day Care and Home Care Providers	Texas	Operation Name, Programs Provided, Accepts Child Care Subsidies, Hours of Operation, Treatment Services, Licensed to Serve Ages

Food Assistance

This data consists of mapping food assets and food subsidy distribution programs, including information on uptake of SNAP programs, and locations of local farmers' markets.

Timeliness: Monthly

Format: Most data on food assets is best shared as geographic information to help people find access points; raw data on SNAP program uptake may also be useful in tabular formats.

Risk level: Medium. Most data on food assets and retailers is non-sensitive, but any data on SNAP recipients must be masked or aggregated to avoid exposing individuals at granular geographic levels.

Connection to federal programs: The U.S. Department of Agriculture (USDA) has an open data policy that supports the publication of farmers market directories or SNAP retail locators.

Examples:

Dataset	State	Sample Fields
SNAP Individuals Enrolled and Dollars	Pennsylvania	Month, County, SNAP Individuals, SNAP Dollars
SNAP Caseloads	New York	District, (Temporary and Non-temporary) SNAP Households, SNAP Benefits, SNAP Persons
Food Assistance by Month	Iowa	Service Area, Households, Recipients, Allotments
Map of Food Access (PA)	Pennsylvania	Address, Distance, Contact Information

Family Assistance

Data about family assistance or general public assistance uptake consists of summary data on requests for any and all safety net benefits for families and individuals.

Timeliness: Monthly

Format: Summary data in tabular formats or visualized formats allows users to understand demographic, temporal, or geographic trends in the distribution of safety net benefits.

Risk level: Medium-high. While summary data is highly useful for policymakers and members of the general public, re-identification risk is high when data is not masked or aggregated to protect the identities of benefits recipients. Geographic masking or point randomization may help to reduce the risk of identifying benefits recipients.

Connection to federal programs: Because public assistance can include various types of safety net programs state-by-state, federal program reporting can vary based on the type of assistance. In most cases, data privacy protections do not apply to summary-level open data about public assistance requests, so states must develop their own data protection practices.

Examples:

Dataset	State	Sample Fields
Family Assistance Cases by Month	Pennsylvania	Address, Total Funded Family Slots, Yearly Goal Number of Families, Cumulative Families Served
Family Support/TANF Program	Iowa	County, Cases, Recipients, Grants
Health and Human Services Program Dashboard	California	Population Below Federal Poverty Level, Education, Demographics, Program Individuals, Participation by Program, Legislative Rosters, Facilities
Total Public Assistance Cases Opened by Month	New York	Temporary Assistance Cases, Recipients, and Expenditures; Family Assistance Federally Participating Cases; Safety Net Assistance Cases; Federally Non-Participating Cases; Safety Net Assistance Federally Participating Cases
Department of Social Services Application Numbers (Daily Cash/Medical/SNAP)	Connecticut	Cash Assistance Applications, SNAP Applications, Non-MAGI Medical Applications

Evictions

This data comes from courts and includes data on eviction notices, active cases, case outcomes, and any other interventions undertaken by the courts. Some data on evictions may also include data on pre-emptive rental assistance programs.

Timeliness: Weekly

Format: Most available open data on evictions is accessible via search tool but, ideally, eviction data would be available in bulk and in interoperable data standards or formats to allow for linkages across the pipeline of housing instability and evictions.

Risk level: Medium-low. Court data is public record, including identifying information about renters, landlords, and property owners. Some data, like individual-level data that may be used to identify early intervention, should remain private or aggregated to a summary level.

Connection to federal programs: None currently but there are efforts underway to create national data standards and databases.

Examples:

Dataset	State	Sample Fields
Evictions Search Tool	Wisconsin	Case Number, Filing Date, Case Status, County, Name, Date of Birth, Caption/ Description
Scheduled and Executed Evictions	City (New York City)	Court Index Number, Eviction Address, Executed Date, Marshal Name, Property Type, Scheduled Status

Healthcare Claims

This data consists of Medicaid, all-payer claims, payments, enrollments, and enrollee demographics.

Timeliness: Monthly

Format: Summary-level data should be available in tabular formats for bulk download.

Risk level: Medium-high. Individual-level data on Medicaid or other healthcare claims should not be published as open data and is protected by HIPAA. Summary level statistics on the number of people served or other demographic characteristics may be of public benefit when published on a monthly basis.

Connection to federal programs: Medicaid data is collected for reporting purposes and can easily be published at summary levels for public benefit. The Department of Health and Human Services (HHS) aggregates and publishes data through the Centers for Medicare and Medicaid Services and creates open data tools like the CMS Mapping Medicare Disparities tool.

Examples:

Dataset	State	Sample Fields
Medicaid Payment and Claims	Iowa	Recipients, Claims, Units, Payments, Average Unit Cost, Average Unit Per Recipient, Average Cost Per Recipient
Medicaid Enrollment by Month	New York	Eligibility Year and Month, Medicaid Aid Category, Enrollment in Medicaid Managed Care or Fee For Service, Plan Name, Number of Recipients
Medical Assistance Enrollment	Pennsylvania	County, Medical Assistance Individuals, Medical Assistance Children

Emergency Shelter Availability

This data contains up-to-date information about capacities at shelters and other temporary or supportive-housing providers.

Timeliness: Daily, and should be consistently updated at a set time.

Format: Data should be easily usable and accessible by case managers, service providers, and individuals seeking shelter.

Risk level: Low. This data does not include any personally identifiable information.

Connection to federal programs: N/A

Examples:

Dataset	State	Sample Fields
COVID-19 Homelessness (Shelters and Hotels Deployed)	California	County, Rooms, Rooms Occupied, Trailers Requested, Trailers Delivered

Mental or Behavioral Health

This data includes statistics about general public health, mental health, and behavioral health services and outcomes. This may also include environmental health as part of general public health reporting, where relevant.

Timeliness: Varies

Format: Most examples of this data are available by visualization or dashboard to demonstrate programmatic or social trends over time, but statistical data on outcomes may also be useful in tabular formats and should be available for bulk download.

Risk level: Medium-high. Mental health and behavioral health data on individuals receiving provider services are protected by HIPAA, but can be aggregated to a summary level and anonymized before publication. Data on general public health and environmental health trends should be captured at the summary level and avoid identification risk.

Connection to federal programs: The Substance Abuse and Mental Health Services Administration at HHS administers a majority of the relevant block grants that fund mental and behavioral health services. The administration has little relevant guidance or practices for publishing summary data at the state and local levels, but does publish some open data for researchers at the federal level.

Examples:

Dataset	State	Sample Fields
California Mental Health Services Dashboard	California	Diagnosis Code, Race, Age, Sex, Service Description, Delivery System, Discharges
Health Statistics Web Tool	Pennsylvania	Fertility Rate, Communicable Disease, Cause of Death, Cancers, Population, Hospitalization Discharges

Opioid Crisis and Response

Data about the opioid crisis includes reporting on outcomes from policy or programs to curb opioid fatalities and prescription drug monitoring. This data may come from a combination of law enforcement, court, emergency medical services, and health provider sources.

Timeliness: Monthly

Format: Data on the opioid crisis and associated response is often published via subject-specific dashboard to allow users to view the intersections of various indicators related to opioid use. Source data should also be made available for download.

Risk level: Medium. If aggregated to summary levels for reporting on program activities and outcomes or provider locations, there should be no personally identifiable information at risk of exposure.

Connection to federal programs: The Medicaid program tracks where prescriptions funded by Medicaid are provided in relation to opioid overdoses. Otherwise, aside from specific reporting to agencies like the Centers for Disease Control, most reporting or analysis of opioid data happens at the state health agency level.

Examples:

Dataset	State	Sample Fields
Opioids Dashboard	Pennsylvania	Individuals Receiving Medication-Assisted Treatment, Treatment Facilities, ER Visits, Hotline Calls, Doses of Naloxone Administered, Drug Overdose Deaths, Rate of Maternal Opioid Use Disorder
Opioid Data Dashboard	New York	Overdose Deaths, Emergency Department Visits, Prescription Monitoring, Hospital Discharges, Youth Risk Behavior, Drug Use

Conclusion

Publishing open data is an essential part of building the public's trust in government, as transparency allows people to see into the everyday operations of state agencies. Accountability measures let residents know that agencies will respond with changes to policy and data reporting structures when statistics are incorrect or when policy outcomes shift.

States attempting to launch open data programs can consider a range of the questions addressed in this report when developing their strategies. Are the short-term goals of the program to establish fundamental transparency and accountability practices, or are they to address critical programmatic and policy needs? What best practices can the program borrow from other states who have launched portals and published usable, impactful data? How can states work with legislators or legal experts to build out protocols for sharing data responsibly?

The increased attention on states' data as a result of COVID-19 and the outsized role of state governments in emergency response demonstrated that the value of open data goes beyond just establishing openness for the sake of it. Journalists, researchers, advocates, and residents all need access to essential information that sits locked up in filing cabinets or agency harddrives.

In addition to allowing this wide range of users access to public data, open data must be usable and well-documented. Documentation—though it can be unpopular to fund or invest in—is an important cornerstone of a strong state data strategy. Robust metadata and clear data governance workflows can help data become more reusable and interoperable across state agencies, and in the general public.

Additionally, states must consider the variety of accessibility needs of various members of the general population. Often, understanding accessibility and making open data more accessible requires speaking to residents and testing data platforms or tech tools with potential users before deployment. Engaging communications or community engagement functions as part of an open data program can allow data owners to open two-way dialogue with potential data users and ensure continuous improvement of their programs.

This guidebook does not contain every impactful dataset being published by a state government. But it does provide a starting point for states beginning to explore which information their residents and partners might want and need.

Appendix: Resources

OPEN DATA

- "Who's at the Popular Table?" Our analysis found which Open Data the Public Likes." Sunlight Foundation
- "Research: Cities can save time on Record Requests by Doing Open Data Right". Sunlight Foundation
- "Third Wave of Open Data". Open Data Policy Lab
- "Open Data Policy Guidelines". Sunlight Foundation
- "State Open Data Laws and Policies". National Conference of State Legislatures
- "US Open Data Toolkit: Best Practices". The Center for Open Data Enterprise
- "State Open Data Portals". State CDO Network

GENERAL PRIVACY GUIDANCE

- "Mythbusting Confidentiality in Public Contracts". Open Contracting Partnership
- "Open Data Aggregation and Suppression Guidelines". Connecticut Open Data Portal

STUDENT PRIVACY GUIDANCE

- "Summary of the Clery Act". The Clery Center
- "FERPA and the Coronavirus Disease 2019". US Department of Education
- "FERPA and Coronavirus Frequently Asked Questions". US Department of Education
- "Does the HIPAA Privacy Rule apply to an elementary or secondary school?" US Department of Health and Human Services
- "Joint Guidance on the Application of FERPA and HIPAA to Student Health Records". US Department of Health and Human Services

HEALTH PRIVACY GUIDANCE

- "February 2020 Bulletin: HIPAA Privacy and Novel Coronavirus". US Department of Health and Human Services, Office of Civil Rights
- "HIPAA and COVID-19". US Department of Health and Human Services
- "Guidance Regarding Methods for De-Identification". US Department of Health and Human Services
- "Disclosure of COVID-19 Information by Domestic Violence Shelters". The Network for Public Health Law
- "Requirements for Notification of Confirmed and Suspected COVID-19 Cases Among Residents and Staff in Nursing Homes". US Department of Health and Human Services, Centers for Medicare and Medicaid Services