Module 6: Introduction to Open Data

Welcome back!

In earlier modules we progressed from learning to define a problem to learning how to do so with the use of data and evidence.

In fact, we had two modules - one on data analytical thinking, explaining why to use data to define your problem, and another on strategies for how to do so, learning to use data in practice to accelerate your own projects.

Now, we deepen our discussion of data to introduce you to one of the most powerful and important governance innovations of the last decade -- the policy of open government data. As we will see, open data policies are a key enabler for one of the most important sources of information and evidence available to you to use when solving public problems.

By the end of this module, you should be able to:

1. Define open data.
2. Understand how open data policies can be used to make data available for public problem solving.
3. Consider what open data might be available to you to use when defining your problem.

Let's get started!

Edmond Halley, the 17th and early 18th century astronomer who lent his name to Halley's Comet, published an article on annuities in 1693.

His population table was based on data collected for the years 1687–1691 from the city of Breslau (now called Wrocław), by the Protestant pastor of the town, Caspar Neumann. This work is now seen as a major event in the history of demography and the first major work of actuarial science.

It is also a wonderful illustration of the value of sharing data openly. By giving Halley the raw information, Neumann enabled more value and insight to be created than had he kept the data for himself.

Such collaboration is what makes open data truly transformative. The organization or individual that collects and maintains information is not always in the exclusive position to use it well. But opening up and sharing data enables the collaboration of people with diverse skills and talents and insights to work together.

By making data open, you enable others to bring fresh perspectives, insights, and additional resources to your data, and that's when it can become really valuable to you and to others for public problem solving.

So what is “Open Data” specifically? Government has always collected data. It gathers information from companies in its role as regulator, it tracks statistics about the economy and society in its role as a policymaking body, and its collects data from citizens in its role as a provider of public goods and services.

But what distinguishes open data from other types of data is that it is publicly available, can be freely accessed and used, and is capable of being processed by a machine.

That is, to be considered “open”, data must be both technically and legally accessible.

To make it **technically accessible**, data must be available in a form that a computer can access and use.

To be **legally accessible**, data must be licensed in such a way that anyone can use and reuse the information without fee or restriction.

When data is legally and technically open, anyone with the right tools, whether they are the data owner or not, can create sophisticated and useful tools, and conduct analysis across datasets to enable empirical problem-solving and advance both social good and economic growth.

Take the example of Mejora tu Escuela. Created by the Mexico Institute of Competitiveness (IMCO), Mejora Tu Escuela is an online platform that makes government data about Mexico’s schools publicly available. The website provides parents with comparative data so that they can compare their own school’s results to others, thereby empowering them to demand better-quality education for their children. It publishes expenditure data, giving activists, administrators, policymakers,and journalists the means to dig deeper, to spot fraud and corruption, and to advocate for change.

This is exactly what happened in 2014, when a report by IMCO revealed that over 1,400 teachers on public school payroll were supposedly more than a hundred years old (with most having the same birthday) and that many earned more than the president of Mexico.

No, the school board had not discovered Ponce de León’s mythical Fountain of Youth. Rather, the story of Mejora Tu Escuela illustrates how, when government makes information free of charge and readily downloadable in digital form, such open data can solve problems.

In this case, federal authorities had required states to provide information about the condition of schools, payrolls, and other expenditures. But it was civil society activists at IMCO who created the platform to make that information accessible to citizens and who also scrutinized that information, ultimately exposing rampant malfeasance that was previously hidden. Although the government initially prevaricated, claiming clerical error, the ensuing media frenzy over the website helped to prompt reform and a shift of responsibility over education from states to the federal government. Ultimately, the activists and the federal bureaucracy worked in parallel, addressing this local-level corruption and acting to improve Mexico’s schools.

Open data matters for the reasons using data matters - we can use it to spot mistakes, outliers, and rare events, and to help us target scarce resources more effectively.

Let's consider a few more examples of open data being put to work.

First, open data sometimes achieves greater government accountability. In the United States, at the federal level, open data facilitated the creation of USASpending.gov, a set of online tools for exploring the federal budget.

Opening local government data about public works in Zanesville, Ohio revealed a fifty-year pattern of discriminatory water service provision. While access to clean water from the City of Zanesville water line spread throughout the rest of Muskingum County, residents of the predominantly African-American area of Zanesville, Ohio were only able to use contaminated rainwater or to drive to the nearest water tower and truck water back to their homes. Opening the data laid the truth bare and led to a successful civil rights lawsuit against Zanesville in 2008.

Second, open data can improve the delivery of services. At the state and local level, increasing access to open data has allowed entrepreneurs and developers to build tools such as smart transit apps, citizen-facing information services, and business- or government-facing data visualization and analysis platforms. For example, both transit authorities and commercial providers use open transportation data to tell commuters when to expect a bus along their route. Retroficiency analyzes energy consumption data to allow utilities, energy service providers, and building owners to identify buildings with high energy savings potential.

Third, open data also enables the creation of tools to improve consumer choice and citizen decision-making in the marketplace. For example, data collected by the government from universities has been transformed by the Department of Education into a calculator—the College Scorecard—to help parents and students make more informed financial decisions about their college education.

Sometimes the benefits of open data ripple out beyond government accountability. For instance, open data can catalyze greater business competition and entrepreneurship. Think of the wealth and jobs created by government's release of both weather data and geo-locational data, which enabled weather apps and GPS devices, respectively. The Open Data Institute notes that the global market for open data could be as high as $5 trillion.

Thousands of companies worldwide now use open government data as a core business asset. One example of this is BrightScope, which worked with previously "locked up" Department of Labor Form 5500 retirement plan data to offer better decision-making tools to investors.

A decade ago, open data was but an idea – a call for action by pro-democracy activists wanting government to be more transparent. Today, it encompasses a broader movement that is focused on solving public problems. Open data policies have helped to drive that change.

On his first day in office in 2009, fulfilling an earlier campaign promise, President Obama signed the Memorandum on Transparency and Open Government, declaring that "[i]nformation maintained by the Federal Government is a national asset," and calling for the use of "new technologies to put information about [agency] operations and decisions online and [to make it] readily available to the public."

In addition, the policy made clear that because the collection of data by government is paid for by the taxpayer, it makes sense to give that data back to the public to use for free.

When the federal government's open data repository, called Data.gov, launched in May 2009, it made forty-seven datasets searchable, turning the principles of the Memorandum into practice by creating a tangible and central place for agencies to list government data and for the public to find it.

Later that year, the Office of Management and Budget directed federal agencies to release not only data about the workings of government but also "high-value" information.

The choice to broaden the forty-year-old definition of government transparency responded to what both the technologies of big data and the technologies of collaboration could make possible today. The directive emphasized the broad public benefits and the need to disclose new kinds of government information as open data, such as locations of reported crimes, weather information, and information that could foster new businesses.

In 2013, the federal government recommitted to its open data policy by issuing an Executive Order on "Making Open and Machine Readable the New Default for Government Information" to advance and accelerate open data implementation in federal agencies. Entrepreneurship and innovation—rather than accountability—are emphasized in the Order. It makes it clear that "making information resources easy to find, accessible, and usable can fuel entrepreneurship, innovation, and scientific discovery that improves Americans' lives and contributes significantly to job creation."

Further laws have followed, broadening the scope of data covered under open data statutes and policies. The Digital Accountability and Transparency Act (DATA) of 2014 calls for publishing all federal government spending data as open data in standardized formats.

There is also the "Open, Public, Electronic, and Necessary Government Data Act" or the "OPEN Government Data Act," which was signed into law in early 2019. It calls for inventorying and publishing all government information as open data.

Today, there are a quarter-million federal data sets online on data.gov, and just about every state and hundreds of cities release some data as open data and have some form of open data portal.

Despite this, the need for continued open data policymaking is as strong as ever. An Open Data Barometer survey of 1,725 datasets covering 115 countries found that nearly 90% of priority datasets remained closed. Only 7% of the data governments collect is fully open, only one of every two datasets is machine readable and only one in four datasets has an open license.

The bipartisan interest in evidence-based approaches to governing has fueled demand for more access to administrative information of all kinds, including the data that agencies collect about companies, workplaces, and the environment.

Using open data is a great way to get data to use to define and understand your problem.

However, before you push ahead to identify how to use open data to better define your problem, remember to make sure you have started by defining the problem as discussed previously.

Without knowing the problem, it will be hard to know what type of data you need so go over the exercises for defining your problem.

Also review the exercises in module 5, designed to get you thinking about what data you need.

But now let's finish by considering whether the data you need is or might be made available as open data.

First, consider the availability of the required data. Is it likely that the government – your own agency or another -- collects the data that would be useful for solving your specific problem?

While you can start with your own state's open data catalog, often these are not comprehensive sources of available data and other relevant agencies need to be engaged to identify available data sets.

There are numerous aggregators of open data you can consider:

For example, try

- US Census Bureau
- Urban Institute
- Depending on your field of interest, different federal agencies such as the EPA for Environmental Data and the US Department of Labor for labor data or the FBI for crime statistics, for example, offer access to free and machine readable data.
- OpenCorporates is largest open database of companies in the world.

Once you find the data, is that data fully open and/or accessible to you in a machine-readable form enabling you to use it readily for analysis?

If not, can you identify external or internal partners with the relevant expertise to help you prepare the data for use. One strategy is to organize a hackathon or datathon also known as a data-dive. DataDives are high energy, marathon-style events where teams of volunteer data scientists, developers, and designers help mission-driven organizations such as government agencies to organize, manipulate, clean or visualize their data.

If the data is not collected, what would it take to collect the data? We have previously discussed methods such as interviews and surveys. And in our next module on open innovation, we will look at how to use crowdsourcing to collect data using distributed participation.

Next, consider your level of readiness to make use of the data. Do you have access to the necessary expertise? Again, reaching out to partners, especially in universities can be one way to obtain the necessary expertise.

Another way is the use of competitions. New York City, for example, gets people to use its data by hosting competitions to attract data-savvy individuals to analyze the data. The City's BigApps

competition invites private companies to solve public problems using open data. The challenge, overseen by the city's Economic Development Corporation (NYCEDC), engages agency leadership throughout the planning process to open up more data. For example, a past BigApps winning team used targeted geolocated data to create Mind My Business, to assist brick-and-mortar food service establishments by sending alerts that help owners predict changes in customer traffic, operate more efficiently, and avoid fines. Private sector platforms like Kaggle offer a community of data scientists online ready to solve problems usually in exchange for a prize.

That concludes our short session on the power of open data. Used well open data can generate new insights and enable us to define problems using empirical evidence. But collecting that data, deriving insight from it and ultimately designing solutions to public problems, require collaboration. In our next module we turn to exploring ways of using new technology to organize such collaboration efficiently and effectively.