Learning to Trust Open Data

By Gadi Ben-Yehuda

Joel Gurin recently released a book enthusiastically titled *Open Data Now*. Gurin, the former chief of consumer and governmental affairs for the Federal Communications Commission, joins a growing chorus calling on the federal government to live up to the spirit of President Obama's 2009 Memorandum on Transparency and Open Government.

Champions of open data exist both within government—Mr. Gurin and the Department of the Treasury's Marcel Jemio, for example—and within industry, including organizations like Socrata, 1776, and XBRL.US. They note that opening government data directly spurs economic activity, enables services Americans depend on every day, and increases the efficiency of and trust in government.

But when vast stores of data are already "open"—accessible to the public, machine-readable, and in a non-proprietary format—what are the next big steps for open data advocates? One obvious step is opening ever-greater troves of data and switching government data's default setting from closed to open. Another is improving the quality of data already available, most notably by ensuring the availability and quality of metadata.

Why Open Data?

Perhaps the biggest success of open data was achieved by accident, and scarcely a panel can be convened or article written without referencing it. In the 1980s, the United States launched satellites into space so the military could have precise location data for training, monitoring, and missions. Nearly two decades later, ordinary citizens were given access to that data stream. The $26.67 billion GPS market is possible only because of that open location-data stream.

There are other examples of open data spurring economic activity. Health data released from the Department of Health and Human Services (HHS) is already powering apps, and HHS regularly participates in "Health Datapaloozas" to bring its data to private-sector developers. Data from the National Oceanic and Atmospheric Administration undergirds almost all weather apps on the market. The Department of Labor publishes data enabling an app that helps with financial decisions.

Economics, important as they are, represent only one part of the story. Another part is the trust in government essential to a democracy. Opening the government's data means everyone benefits from their government. Everyone becomes a stakeholder and sees the value they personally derive from their government's activities. And opening the data about how government operates allows everyone to understand how public money is spent and see the alignment between public priorities and public expenditures.

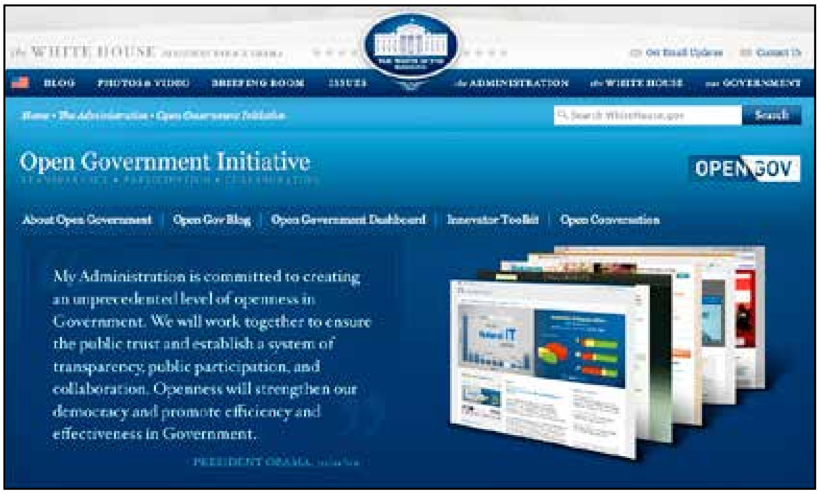
Numbers Don't Lie

The popular saying is that "numbers don't lie," but it can be countered with the equally popular "lies, damned lies, and statistics." When it comes to big numbers, this is even more true, as humans are famously bad at grasping the meaning of large numbers.

Gadi Ben-Yehuda is the Innovation and Social Media Director for the IBM Center for The Business of Government.

Few likely know that better than Earl Devaney, a former inspector general for the Department of the Interior and chairman of the Recovery Accountability and Transparency Board. Mr. Devaney was asked by a congressional oversight panel to estimate how much Recovery Act money would be lost to fraud, waste, and abuse. A 2009 study found that those losses typically consumed between five and seven percent of a government program's budget. While that may not sound like much, the Recovery Act had a budget of $787 billion, which grew to $831 billion through subsequent legislation. So the raw number for waste? Between $40 and $55 billion projected to be lost. Both numbers are accurate, but each tells a different story.

What makes the Recovery Act such a good example is not the amount of money it was projected to lose, but the amount of money it did lose. Mr. Devaney writes in *Fast Government*, "The remarkable success the [operations center] has had in minimizing fraud and waste is evidenced by the numbers: Less than one-half of one percent of the nearly 277,000 contracts, grants, and loans awarded under the Recovery Act are under investigation. This pales in comparison to the five-to-seven percent figure normally associated with losses for any large government program."



My Administration is committed to creating an unprecedented level of openness in Government. We will work together to ensure the public trust and establish a system of transparency, public participation, and collaboration. Openness will strengthen our democracy and promote efficiency and effectiveness in Government.

PRESIDENT OBAMA

And the important difference between this program and most others was that the financial data for the Recovery Act was designed to be open from the start. The GPS industry and the Recovery Board examples speak to the first goal of open data advocates: opening more stores of data. How many industries are simply waiting for businesses large and small? How much more effective will current industries and markets be when they have access to data that is currently inaccessible to them? Further, open data advocates point to the increased efficiencies that could be realized if more people had access to more data.

And "more data" is where the proponents of metadata find common cause with their data-set-oriented comrades.

The Importance of Metadata

Marcel Jemio, the chief data architect in the Department of the Treasury's Bureau of the Fiscal Service, is a cheerleader for metadata. He uses the metaphor of apple varieties (discussed below) to illustrate the value of metadata. He says that from metadata, people can derive context, understanding, quality, security, analytics, worth, trust, and ultimately, innovation.

To understand the importance of metadata, think of a digital photograph with the caption "Sun Rising over Miami Beach." The metadata for digital photographs is called "EXIF" and it has certain attributes: the kind of camera that captured the image, the time it was taken, the f-stop and aperture, whether a flash was used, sometimes even the geolocation. If, looking at the EXIF metadata, one saw that the picture was taken at 8:00 PM with a camera located 20 miles east of Miami (that is: from a boat), one would know that it was not sunrise at all, but sunset. The photographer's veracity would be called into question, and their other work would be subjected to further scrutiny. This is why Mr. Jemio is right to say that metadata can give context (it is sunset, not sunrise), and trust (in the form of verifiability).

There are other examples of metadata adding value to data sets. One company that puts EXIF metadata to fascinating—and meaningful—use is OKCupid. In a 2010 blog post titled "Don't Be Ugly By Accident!," the site's data analysts "aggregate[d]11.4 million opinions on what makes a great photo." They then analyzed the responses and determined which brands of cameras took the best pictures, what time of day was optimal, what f-stop made people look more attractive, and how the use of flash was likely to return a better picture. This analysis was performed not using the data—the image—but using metadata. And with that analysis, people could create better data; that is, they could take better pictures!

It is easy to extrapolate meaningful government uses from this. Metadata can accompany any data. Take produce, specifically apples. While famously not comparable to oranges, apples seem like they should be comparable to one another, yet there are many varieties of apples and even within those are regional differences and other distinguishing qualities that describe a specific fruit. These metadata give context, allow for analysis, instill trust, provide specificity, and most important, make it more likely that people can use the data in ways that add value both for themselves and for the larger economy.

Why the Future Is Open

Two developments point to a bright future for open data advocates. The first is the proliferation of tracking devices and software in every facet of American society. The complementary development is the growing sophistication in understanding both raw data and the visualizations built on that data.

Data trackers are quietly moving into every part of our lives: "Automatic" is a device that plugs into a car's computer and relays real-time data about fuel efficiency, engine operations, and vehicle location. Body trackers have gone mainstream, and more people are counting their steps, monitoring their heartbeats, and using WiFi scales to see not only weight, but body composition. Even school report cards are using data visualization, not simply reporting raw data in the form of letter grades or percentages. And as people grow accustomed to seeing data in all parts of their lives and appreciate how it is helping them make better decisions, they will press for open data from their government.

Appropriately, the public is also learning how to interpret data with more nuance and sophistication. One concern about examining and releasing data is what it will reveal. People and organizations don't always accomplish their goals, and when they do, it may be with some degree of waste or inefficiency. But performance is increasingly seen through the lens of data visualizations and dashboards, and people can see that sometimes they do not meet all their targets. They also see that success is often a sliding scale, not a threshold to be crossed.

All this points to a future in which more people will clamor for data and there will be less concern about releasing it. And as the government accedes to the requests for more and better data, both the government and the citizens it serves will be better off. ◘