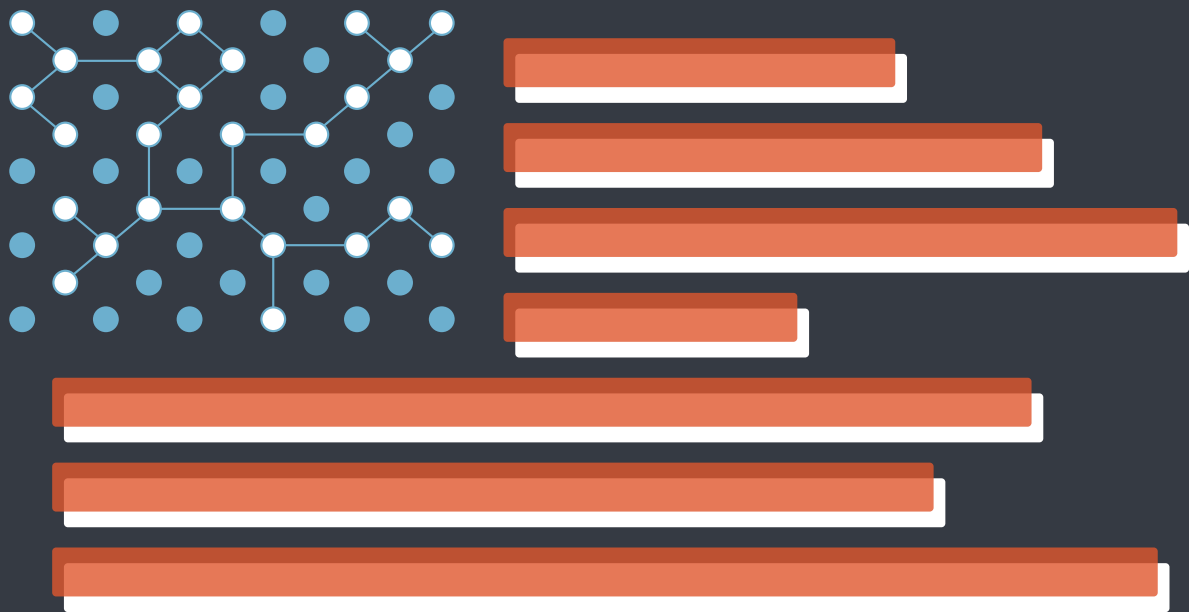# THE
STATE OF THE UNION
OF OPEN DATA, 2016

THE
STATE OF THE UNION
OF OPEN DATA, 2016

by Alison Gill, Hudson Hollister, and Adam Hughes

DATA FOUNDATION

The Data Foundation is the nation's first industry-focused open data research organization. Through research, education, and programming, the Data Foundation supports the publication of government information as standardized, open data.

http://www.datafoundation.org/

GRANT THORNTON PUBLIC SECTOR

Grant Thornton Public Sector helps executives and managers at all levels of government maximize their performance and efficiency in the face of ever tightening budgets and increased demand for services. We give clients creative, cost-effective solutions that enhance their acquisition, financial, human capital, information technology, data analytics, and performance management.

https://www.grantthornton.com/publicsector

INTRODUCTION

Data Transparency 2016, the fourth annual policy conference hosted by the Data Foundation in September 2016, was a significant step forward for the open data movement in the United States. Over 1,000 government, nonprofit, and industry leaders came together to learn about developments in the field, share best practices, and work toward a future in which more open government data will bring better management and oversight, increased private innovation and competition, and a better basis for shared decision-making. *Data Transparency 2016* occurred at a historic time for the open data movement, with the Digital Accountability and Transparency Act (DATA Act) set to take effect in May 2017, requiring every U.S. federal agency to begin reporting its spending as open data. The DATA Act will transform the federal government's spending information into a single, standardized open data set – perhaps the most valuable government data resource in the world.

The U.S. open data movement has no leader, but is invigorated by the nation's government, nonprofit, and tech-industry sectors. Because *Data Transparency 2016* brought together all of the major constituencies of the open data movement, we believe the combined perspectives of its participants represent the *State of the Union* of open data in the United States, in a way that one leader's views never could.

Accordingly, we conducted interviews with more than 40 presenters at *Data Transparency 2016*, including Congressional and agency leaders, open data experts and advocates, and private sector leaders, and we requested written submissions from many others. By combining their perspectives, this report seeks to capture the moment, provide a vision of the future, and catalyze further efforts for the open data movement.

Even among *Data Transparency 2016* presenters, opinions of the significance of open data ranged from those calling the movement a "radical transformation" of our society and a "monumental change" to those calling open data a "buzzword" or "incremental at best." This much is certain: The U.S. federal government is committed to releasing unprecedented levels of federal spending data going forward; proposed open data policy reforms enjoy strong bipartisan support; and federal agencies are more than ever before making standardized data available to the public. This report seeks to illuminate the areas of agreement and show how they might shape the future of the open data movement – and the future of U.S. government and society.

🎤 INTERVIEWS WITH PRESENTERS

Each conference presenter was asked :

1. What does the term 'open data' mean to you?

2. What types of government data do you most use in your work?

3. The Data Foundation tends to define open data using two steps: first, standardize; and second, publish. First, what role do data standards play in your work and your industry?

4. Second, what role do the publication and free availability of data play in your work and your industry?

5. Have the standardization (or comparability) and the publication (or availability) of data related to your work improved in the last five years? In what ways?

6. What do you think will be the most important outcome of more open data in your field? A) Lower reporting costs; B) Ability to cross-compare data; C) Ability to provide greater predictive analysis; D) Greater availability of data to various constituencies; E) Other

7. What is your vision for the future of open data and data transparency?

OPEN GOVERNMENT DATA PRINCIPLES

According to the Sebastopol Group, December 2007

1. **Complete** – All public data is made available. Public data is data that is not subject to valid privacy, security, or privilege limitations.

2. **Primary** – Data is as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms.

3. **Timely** – Data is made available as quickly as necessary to preserve the value of the data.

4. **Accessible** – Data is available to the widest range of users for the widest range of purposes.

5. **Machine processable** – Data is reasonably structured to allow automated processing.

6. **Non-discriminatory** – Data is available to anyone, with no requirement of registration.

7. **Non-proprietary** – Data is available in a format over which no entity has exclusive control.

8. **License-free** – Data is not subject to any copyright, patent, trademark, or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed.

WHAT IS OPEN DATA?

What do you think of when we use the term 'open data'? Do you think about the multi-billion dollar transportation companies which rely upon Global Positioning System (GPS) data to function? Do you think of the atmospheric data which allows us to forecast the weather, prepare for potential disasters, and plan agriculture? Do you think about the information released through leaks and whistleblowers which, for better or worse, affects the way we think about government programs and our leaders? Do you think about the data analysis which has allowed investigators to uncover fraud at major financial institutions? As you can guess, the availability and use of data has a major impact on each of our lives.

Simply put, **data**[1] is electronically stored information. There are a multitude of examples of data held by individuals (an online friends list), companies (phone, medical, and banking records), state and local governments (property tax and spending information), and the federal government (crime statistics, survey data, Medicare data, spending data). The focus of *Data Transparency 2016* and of this report is U.S. government data, especially federal data.

The concept of **data transparency** refers to the ability to easily access and work with data regardless of how the data was originally recorded or where it is located. **Machine readable** data is expressed in a format that a computer can process. **Data standards** are rules by which data is described and recorded, allowing users and their software to automatically understand meanings and relationships. Examples include the Project Open Data Metadata Schema, which ensures all of the data sets published on the U.S. government's Data.gov portal are identified and described consistently, and the U.S. GAAP Taxonomy, which organizes the electronic version of U.S. public companies' financial statements prepared using Generally Accepted Accounting Principles. **Data formats** are used to express and apply data standards. For example, the Project Open Data Metadata Schema is expressed using JSON (JavaScript Online Notation), while the U.S. GAAP Taxonomy is expressed using XBRL (the eXtensible Business Reporting Language).

---

[1] 'Data' is the plural of the word 'datum,' but is increasingly used as a singular noun.

Open data is a more complex and subjective concept. In 2007 a group of advocates gathered in Sebastopol, California, to create a set of eight principles[2] that are still considered the first comprehensive attempt to define what open data means in government (see sidebar).
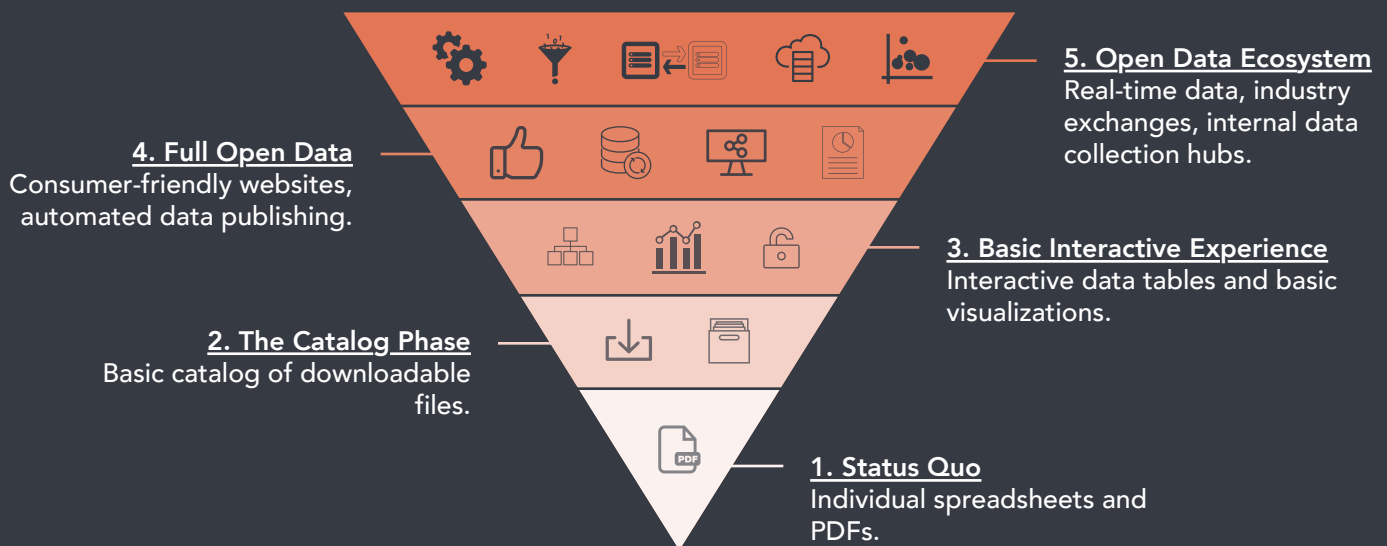
The Data Foundation takes a simpler, even simplistic, view: open data encompasses two broad steps, Standardize and Publish. *Standardization* is the critical process of developing and implementing data standards based on a consensus between creators and users of data. *Publication* means data sets should be available online, without licensing restrictions, and easily downloadable in bulk.

Data Transparency 2016 speakers mostly agreed with the Data Foundation's two-step summary, but many added one or two additional criteria, often timeliness and completeness (see sidebar). Meanwhile, commercial open data vendors have gone beyond defining open data to identifying steps in its deployment and use (see sidebar).

---

[2] https://opengovdata.org/

OPEN DATA MATURITY MODEL
According to Socrata, Inc.

5. Open Data Ecosystem
Real-time data, industry exchanges, internal data collection hubs.

4. Full Open Data
Consumer-friendly websites, automated data publishing.

3. Basic Interactive Experience
Interactive data tables and basic visualizations.

2. The Catalog Phase
Basic catalog of downloadable files.

1. Status Quo
Individual spreadsheets and PDFs.

The board of directors of the Data Foundation gathered at Data Transparency 2016 to celebrate the promise of open data for government and society. From left to right: John Runyan, Runyan Public Affairs; Frank Landefeld, MorganFranklin; Vasan Thatham, DataTracks; Demek Adams, Grant Thornton; Bryce Pippert, Booz Allen Hamilton; Adam Roth, StreamLink Software; Hudson Hollister, interim president of the Data Foundation; Kevin Merritt, Socrata; Sherry Weir, Kearney & Co.; Mike Starr, Workiva; Jonathan Elliott, RDG Filings; Joe Kull, PwC. Not pictured: Craig Clay, Donnelley Financial Solutions.

WHY OPEN DATA?

Open government data is a powerful resource that our society has only just begun to harness. Access to public-sector information as open data can of course tell us more about what government is doing – how money is spent, how programs are performing, whether progress is being made to address persistent societal issues – but this is just the smallest fraction of the immeasurable economic value of open data.

Within government, open data greatly reduces the costs of sharing and using information. When information is expressed as open data, a range of functions become easier and cheaper: measuring program performance, discovering fraud and waste, improving citizens' customer experience, and informing decisions. Many such changes may begin at the federal level and among leading states, but the advantages will flow to state and local governments as standards are created and publication becomes cheaper and more efficient. Other efforts, like the Open311 standard for tracking location-based government services, began at the state and local level and were later adopted at the federal level.

In the private sector, open data can help investors better understand risk and opportunity and provide communities with information to advocate for change and improve their lives. Moreover, the standardization required for open data allows software to automate formerly-manual compliance and reporting tasks. For example, when the Internal Revenue Service and the state tax agencies adopted a consistent data standard for tax returns in 2001, they enabled software packages such as TurboTax to greatly reduce taxpayers' time and cost relative to the complexity of the U.S. tax code. The Dutch and Australian governments have gone much further, adopting data standards across multiple agencies' reporting requirements to allow companies to automatically comply with multiple regimes simultaneously.

Just as the publicly-available data on our genetic code has fostered a rapid growth of knowledge in biology and medicine, and the communication standards which allow for the Internet have affected nearly every aspect of our lives, open government data has the potential to dramatically impact our civil society.

We asked our interviewees about their understanding of open data. Here are some of their responses:

AMY DEORA
Director of Government Analytics at Civis Analytics

To me, the 'Open Data' movement means that data are treated as the valuable resource they are, with the default being that all data created by government would be shared unless there are specific privacy or security concerns.

REED WALLER
Assistant Director of the Financial Management Initiatives Group at the U.S. Department of Justice

Open data means that information is made available to the people on whose behalf it has been gathered. For instance, information that governments gather is made available to the public, on whose behalf government acts.

GARY BASS
Executive Director of the Bauman Foundation

Open data is a subset of open government and consists of two components, 1) technology to make data available and 2) policies that make the right info available to the right people.

PERSPECTIVE: DEMOCRATIZING DATA

Kevin Merritt, founder and CEO of Socrata

Government agencies tend to be cautious about trying new things. Public servants have the responsibility to be prudent when they spend taxpayer dollars and even small changes to a government program can affect thousands and sometimes millions of people. Until recently, the risks associated with change were too great, but today, several federal agencies are embracing technology and data in new ways to make government more responsive to people's needs. They're doing it by making information that was previously difficult to find and interpret more widely available to the public. Think of it as democratizing data.

Two years ago, the Digital Accountability and Transparency Act of 2014, known as the DATA Act, tasked the federal government with transforming spending information into open data. The legislation recognized the value of machine-readable data to do things that were impossible with paper documents and other antiquated reporting systems.

While the DATA Act's full implementation won't happen overnight—it hits the executive branch in May 2017—agencies are applying this approach to other valuable information resources and realizing the benefits of democratized data. Consider the Centers for Medicare and Medicaid Services (CMS), which is required as part of the Affordable Care Act to collect and display information about pharmaceutical companies' gifts and payments to physicians. CMS now freely shares this valuable data with the public on the CMS Open Payments website. The site provides simple tools to search for a doctor or teaching hospital that has received payments, or for a company that has provided payments.

This level of transparency was unprecedented when the site launched in 2015. Today, users can explore and access more than 28 million records to better understand the nearly $17 billion in payments made between 2013 and 2015 in order to make more informed healthcare decisions.

Other agencies, such as the US Agency for International Development (USAID) and the Federal Communications Commission (FCC), are freely sharing data both to improve the efficiency and management of their programs and to provide the public with better information for decision-making.

While several agencies are taking steps to democratize data, many federal agencies and programs are still using paper and other archaic systems to collect information. It will take broad support to transform government information from disconnected documents into open data. That's why I support the Data Foundation, whose research and education help inform the open data transformation.

A BRIEF HISTORY OF THE OPEN DATA MOVEMENT

The concept of open data has historical roots that stretch back to our nation's founding. Indeed, the Constitution, as part of the framework of our system of government, established a Census to be conducted every ten years in order to base representation and governance on actual population numbers. This idea built on earlier, Enlightenment-Age philosophies about democracy, the role of government, and the ultimate power of the people. Based on such an understanding, data about the people and gathered on behalf of the people ultimately belongs to the people, and the government has a responsibility to act as a good steward for such data. The Founders demonstrated a high regard for the open flow of information through the free speech and open press protections of the First Amendment. Noting the critical importance of an informed citizenry to democracy, Benjamin Franklin proclaimed, "Whoever would overthrow the liberty of a nation must begin by subduing the freeness of speech."

Throughout U.S. history, we have seen government information used by not only government officials, but by industry and reformers to improve American society. One early example is the pioneering work of Jane Addams, a Nobel Peace Prize awardee who began her work in the late 19th Century and is widely recognized as the founder of the social work profession. Using new techniques like statistical mapping, Ms. Addams helped build more sophisticated understanding of social ills such as typhoid, infant mortality, and truancy, and to address such issues with targeted solutions.

In the latter half of the 20th Century, the nation saw advances in the legal requirements for openness, and eventually open data, as technology improved. Congress passed the Freedom of Information Act (FOIA) in 1966, giving the public the right to request records from federal agencies. The law greatly expanded the public's ability to access federal information, but created a default system in which government information remains private unless requested, rather than one where all data is public unless there is a specific reason to withhold. Moreover, FOIA is limited in that the requester must know that information exists in order to request it. FOIA does not require agencies to conduct research, analyze data, or create new records; data can be requested in an electronic format, but that does not mean that the data are open or machine readable; and data is only provided to the requester and not to the public at large.

The roots of open data in democratic philosophy were reflected by several of our interviewees:

REPRESENTATIVE SETH MOULTON
(D-MA)

Investing in civic technology and committing to government transparency are key to a modern democracy. In 2016, citizens expect and deserve access to information about the laws that impact them on a daily basis.

SEAMUS KRAFT
Executive Director of the OpenGov Foundation

Improved access to legal materials will allow us to better solve our shared challenges and to better govern.

TERRY GRAFENSTINE
Inspector General of the U.S. House of Representatives

Informed citizens can understand bad choices, make better choices, and hold leaders accountable.

A TIMELINE OF OPEN DATA

1788 Constitution ratified, requiring a Census every 10 years

1807 National Geodetic Survey established to define a national coordinate system

1870 National Weather Service established to provide weather forecasts

1966 Freedom of Information Act passed

1989 Right to Know Network (RTK Net) established

1994 Carl Malamud launches public version of the EDGAR database for SEC filings

1996 Single Audit Act Amendments enacted

2004 Microsoft files first open data SEC filing using U.S. GAAP Taxonomy

2005 National Information Exchange Model (NIEM) founded

2006 Sunlight Foundation founded

2006 Federal Funding Accountability and Transparency Act (FFATA) enacted

2007 Open Government Data Principles developed in Sebastopol, California

2007 USASpending.gov launched

2009 President Obama issues Open Government Directive

2009 SEC begins phasing in mandatory open data reporting using U.S. GAAP Taxonomy

2009 Data.gov launched

2010 Uniform Open311 adopted at the federal level

2011 FOIA.gov launched

2012 Data Transparency Coalition (later called the Data Coalition) founded

2013 President Obama issues Executive Order 13642, Making Open and Machine Readable the New Default for Government Information

2014 Digital Accountability and Transparency Act (DATA Act) enacted

2014 Joel Gurin publishes Open Data Now

2016 Data Foundation founded

2016 Data Transparency 2016 conference

2016 U.S. Data Federation initiative launched

2017 DATA Act requires publication of federal spending data in standardized format

With appreciation to Socrata, Inc.

One of the most important developments in the field occurred in 1994, when Carl Malamud, often considered the modern open data movement's founding father, launched a free, online version of the Securities and Exchange Commission's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system, which publishes documents filed by public companies and financial entities. The SEC had previously charged users for access. Malamud's platform rapidly gained more users than the official one, prompting the SEC to eventually open up EDGAR's contents – today more than 21 million filings – for free public access.

In 1996, Congress passed the Single Audit Act Amendments, which simplified audit reporting for federal grantees. Grantees' audit documents are now publicly available through the Federal Audit Clearinghouse. The vast majority of Federal grants are awarded to state and municipalities, so the system contains significant financial information on local governments. Over time, the system has been updated to require all submissions in an unencrypted document format and to provide full audits packages rather than just information sheets.

In 2006, Congress passed the Federal Funding Accountability and Transparency Act (FFATA), championed by then-Senator Barack Obama and Senator Tom Coburn, which required the White House Office of Management and Budget to publish information on all federal grants and contracts.

The Single Audit Act, FFATA, and similar reforms mandated for government information to be published, but did not accommodate the need for consistent standardization. But after the Sebastopol principles were published, data standards became a concern for reformers.

The Obama Administration demonstrated a strong interest in open data. On President Obama's first day in office, the White House issued a memorandum to require agencies to expand public access to federal data by publishing high-priority data sets in open formats. Following this directive, new websites were created to improve public access to machine-readable data sets, including Data.gov. Since its creation in 2009, more than 200,000 federal government data sets have been published through Data.gov, which is maintained by the General Services Administration. Data.gov uses a standardized format to track the characteristics of each data set, but the data sets themselves need not follow any comprehensive standards. To begin addressing the need for standardization, in 2016, the GSA began the U.S. Data Federation, an initiative to highlight available data standards and encourage agencies to use them.

On May 9, 2013, President Obama signed Executive Order 13642, Making Open and Machine Readable the New Default for Government Information. The executive order directed all federal agencies and departments to make inventories of their data assets and move toward standardizing and publishing all of them as open data sets.

As the U.S. government's willingness to transform its information into open data grew, so too did practitioners' and advocates' desire to understand its history and implications. Joel Gurin published the most comprehensive treatment of the open data movement in 2014 with his book *Open Data Now.*[3]

Several *Data Transparency 2016* speakers said the motivation for governments to express their information as open data is currently changing from external transparency to internal management. In other words, governments are beginning to treat open data as a resource for their own decision-making, and recognize open data sets as the official versions of information, rather than publishing data sets merely to comply with transparency requirements. With this official use of open data sets, decision-makers rely on the same information, and sometimes even the same systems and software, as their external constituents do.

THE DATA ACT: A TURNING POINT FOR OPEN DATA

Perhaps the most-discussed open data development, both among our interviewees and at *Data Transparency 2016*, was the Digital Accountability and Transparency Act (DATA Act) of 2014, the nation's first open data law. The DATA Act added a strong data standardization mandate to prior FFATA legislation by directing the Treasury Department and the White House Office of Management and Budget (OMB) to establish, and enforce, a consistent, government-wide data structure for federal spending information. The DATA Act also expanded the scope of FFATA's publication mandate to include agencies' financial performance and budget actions, in addition to the grant and contract summaries already being disclosed. By May 2017, every federal agency must begin reporting a fully-standardized version of its financial, budget, grant, and contract information to Treasury and OMB; Treasury and OMB must then publish a single open data set covering the entire executive branch's spending. The Data Foundation and MorganFranklin explored the vision behind the DATA Act in July 2016 by co-publishing *The DATA Act: Vision & Value.*[4]

> "Open data can and should be used to improve the lives of the citizens government programs were designed to assist. Currently, when we evaluate programs, we often collect data anew. Open data offers the promise that we could more easily assess how effective programs are and at what cost."

– Robert Shea, Principal at Grant Thornton

Many of our interviewees predicted the transformation of the federal government's spending information from disconnected, purpose-built databases into a single open data set will deliver transformative public transparency. Others predicted the DATA Act, once standardized reporting begins, will create new applications for internal management. For example, former Postal Service inspector general David Williams pointed to the use of standardized spending data to reduce the preparation time for anti-fraud analytics projects (see Perspective on page 9).

The DATA Act did not specify the particular data standard to be used to report and track federal spending information. Instead, it directed Treasury and OMB to create one. Treasury led a public collaboration process, with preliminary versions of the standard published periodically on a GitHub portal, culminating in April 2016 with the publication of the first complete version the **DATA Act Information Model Schema (DAIMS)**. Our interviewees expressed optimism the DATA Act Information Model Schema can be expanded to standardize even public-sector spending information not covered by the DATA Act, such as state and local governments' finances.

---

[3] www.opendatanow.com/

[4] http://www.datafoundation.org/data-act-vision-and-value-report/

Alongside its mandate for federal agencies to begin reporting standardized spending information, the DATA Act also established a pilot program to test the use of standardized reporting by recipients of federal grants and contracts. This pilot, managed for contractors by OMB and for grantees by the Department of Health and Human Services (HHS), will continue through mid-2017. *Data Transparency 2016* speakers expressed a hope that standardized reporting will improve the transparency of federal grants and contracts to the public; empower new oversight tools for grantor and contracting agencies; and allow software to automate formerly-manual compliance processes, reducing recipients' compliance costs, and ultimately saving taxpayers' money.

The DATA Act is the federal manifestation of a nationwide wave of open spending data. At *Data Transparency 2016*, deputy Ohio treasurer Seth Metcalf demonstrated Ohio's Online Checkbook, which publishes the state's spending information as open data. Ohio's Online Checkbook discloses state spending at the transaction level – a degree of detail the DATA Act does not require for federal spending – and relates each transaction to its budget category, department, and office. The platform's development was eased by the state's implementation, several years earlier, of a statewide financial system, which means all state agencies already followed consistent spending data standards.

Our *Data Transparency 2016* speakers were asked: Have the standardization (or comparability) and publication (or availability) of data related to your work improved in the last five years?

Yes   No   Yes, with limitations
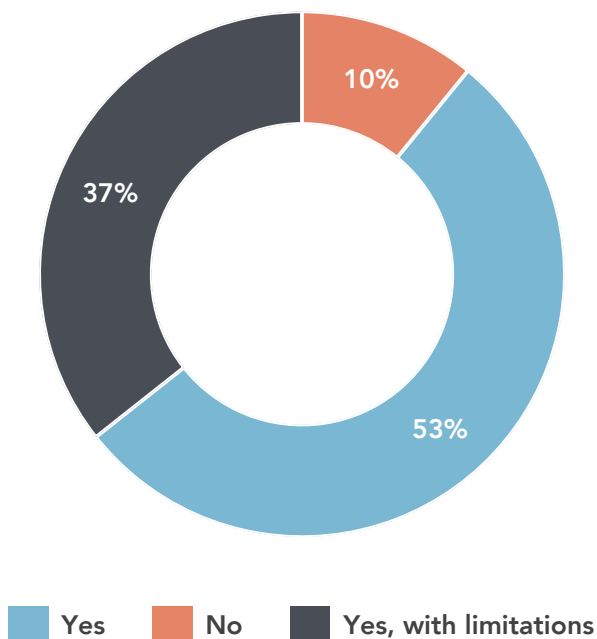
PERSPECTIVE: THE VIENNESE GAMBIT OF DATA

Dave Williams, former Inspector General of the US Postal Service

In chess there is a sequence of moves called the Viennese Gambit. The sequence appears inconsequential, but if not countered immediately, it can't be stopped. That's the DATA Act; it changes everything, forever. Government will be enabled to administer itself within a data-rich environment. Instinct and intuition will give way to evidence based decision-making and evaluation. There will be winners - top performers, wise investments - and there will be losers - poor performers, wasteful programs, and those involved in fraud.

Among the winners, those struggling to conduct data mining will break from the pack. Bad data takes up fully half the time of data analysts. The ETL (Extract, Transform, Load) process is mindless and soul-crushing. That will be greatly simplified. The early entrants into anti-fraud data analytics have experienced some pretty astounding feats already. Well, stick around. It's going to get scary good. Clean and standardized data will increase current efforts by an order of magnitude. All those clever ideas for attacking fraud will suddenly begin breeding.

A feature of fraud data analytics is silent convergence. At the speed of light and very quietly, fraudsters find the police standing next to them as they steal tax dollars. Importantly too, as things get less complex and frustrating, there will be lots of new entrants, who will learn from one another, and from the pioneers, all the tricks of the trade on day one. Federal Inspectors General are already sharing tools and sources of data on a shared website platform called DANTES, first established by my office.

Data Analytics, which sift through big data and operate right on top of organizations' existing data stores, are natural partners with alerts emanating from the Internet of Things, which is also unfolding. Together, leads from data analytics and ongoing crime alerts from the IOT can be aggregated into actionable referrals that might just initiate a Viennese Gambit of their own.

DATA STANDARDIZA-TION

Why are data standards so important to open data that the Data Foundation treats standardization as one of two indispensable steps? The screw is a useful example. Screws have been used to fasten materials together, such as armor and building materials, since the late Middle Ages. They were formed from nails made by blacksmiths by cutting a slot in the head and filing the threading by hand. As you can imagine, this didn't create a great deal of consistency, resulting in poor and uneven threading, incompatible screw sizes, and a very high cost of production. Moreover, early screws could not be used without a drilled hole – and of course, if the hole was larger or smaller than the size of the screw, the fastening would not work. The first screw factory was built in 1760. It took the owner 16 years to raise the necessary capital and even then, the business soon failed.

The screw factory was a step in the right direction – it made screws cheaper and more available. But imagine what might have happened if our first screw factory owners had spoken to the drill factory down the road and agreed to make their various screws and drills in uniform sizes. Think about the screws we have today – they come in well-defined and standardized sizes made for different purposes. Standardization allows many different people in an incredibly wide array of industries to collaborate and make interchangeable parts without even speaking to each other.

For data, the need is similar, even if the tools are different and the scope is much larger. Data.gov alone has more than 200,000 data sets in a wide variety of formats. As with standardized screws, however, the benefits of standardized data are undeniable:

- **Efficiency of production and consumption.** Data sets that are standardized can be more easily produced and used through automated processes and with non-customized tools.
- **Increased comparability.** Standardized data sets can be compared with other data sets for different programs and fields, allowing for new insights.
- **Increased consistency.** Standards can provide well-defined meanings for data elements and help reduce the scope for human error when manual transcription is unavoidable.

- **Greater attractiveness for investment.** Investors are more likely to provide capital if programs and companies that analyze or work with data can be brought to scale through data standardization.

It is for these reasons that the Data Foundation emphasizes the importance of data standards as one of only two key requirements for open data.

With these compelling benefits, as the Treasury Department's Office of Financial Research put it in a 2014 annual report, "*Why have standards not become ubiquitous? Standards are often a classic public good, with costs borne by a few and benefits accruing over time for many.*" That is, there is a collective action problem. Creating standards is a complex process that requires hard work, resources, collaboration, and some risk. For financial data, for example, the process of standardization can mean a minute examination of each financial term reported by a number of entities to determine how the relationships among the data vary and how broader categories are defined, and then setting very precise rules for common data elements and formats. Standards are most often created though market tradition, private sector collaboration, by standards-setting organizations, by government entities, or through some combination of these elements.

Development of standards by industry or standards entities can sometimes lead to standards which are proprietary or which contain intellectual property of one or more of the participants, which can greatly limit the effectiveness of such standards. If our hypothetical screwmaker and drillmaker agreed to set standard sizes but refused to share them with others, it would limit not only limit their use in the field, but it would ultimately limit the growth of their businesses. Therefore, good standards are nonproprietary, widely available, applicable to a variety of contexts, and not too complex. Some *Data Transparency 2016* participants said the federal government must end any use of proprietary data standards, particularly the DUNS Number, which is owned by Dun & Bradstreet, Inc., and universally used to identify federal contractors.

Government can help to solve these issues by mandating standards and setting up processes for their development and maintenance. For example, the Securities and Exchange Commission designated the Financial Accounting Standards Board (FASB), which governs the substantive rules for public companies' financial statements, to maintain the data standard used for the open data versions of those financial statements. Meanwhile, although the General Services Administration does not require data sets published on Data.gov to follow any standard, its new U.S. Data Federation[5] initiative seeks to encourage agencies to voluntarily standardize by highlighting standards that are already available.

---

[5] http://federation.data.gov/

PERSPECTIVE: DATA WITHOUT STANDARDS

Waldo Jaquith, 18F, General Services Administration

The rate at which open data has been adopted by government at all levels exceeds substantially the rate at which the practices and standards relating to open data have matured.

Currently, open data practices are appallingly crude. A list of 100 core types of government data would find that for more than 90, there exists no standard schema. But even generating that list would be a feat, because there's nothing approaching agreement on what the 100 core types of government data are. There's no central data set of government data repositories and their inventories, and no standard way to locate them. Most data sets have no metadata, making it impossible for software to do anything meaningful with them without a lot of manual labor. There is no meaningful adoption of data portability practices, meaning that moving data from one system to another is a multi-hour process. Some crucial data sets are enormous, but must be retrieved in their entirety after even a tiny update, because data set diffing, patching, and segmentation has not been adopted by government at any level. So we also lack data synchronization—the ability to automatically update a local copy of a data set from a remote copy.

There are clever projects working on addressing all of these gaps, and some of these have achieved modest success on the sharing of data within the private sector and academia. But they have not made any impact on the governmental practice of open data, nor is there any sign that they will soon.

An enormous chunk of published government data sets are updated only when a human manually exports data from a system and uploads the data to a repository. The open data ecosystem is less an ecosystem and more a collection of hacks and workarounds, dependent as they are on data otherwise trapped within "enterprise-grade" software built by companies with a financial interest in locking agencies into their multi-million dollar software by keeping the data within from getting out.

The open data movement needs standards, schemas, and infrastructure. As open source has moved into an era of automation and containerization, so too must open data, before its laborious, antiquated practices cause the movement to collapse under its own weight.

We asked our interviewees about the importance of data standards in their fields:

MIKE STARR
Senior Vice President of Government and Regulatory Affairs at Workiva

Organizations can't effectively report data unless it is standardized. They don't see the benefits, can't consume data in an automated process, or analyze data unless there is a standard.

ROBIN CARNAHAN
Director of State and Local Practice at 18F, General Services Administration

Standards are crucial because, otherwise, data remain in silos. Good standards reduce the workload on people and save a lot of time through automation.

BRYCE PIPPERT
Principal at Booz Allen Hamilton

Publishing data is not sufficient for value creation. We need data that can freely move and be understood – and doing that well requires standardization.

DATA PUBLICATION

Due to both legislation and the Obama Administration's policy promoting government transparency, the availability of government data has grown dramatically over the last several years. Existing data sets have been expanded and a vast array of financial, programmatic, and other data sets have been published on new websites. In addition, many agencies and state governments have deployed advanced new tools to allow users to sort and visualize these data sets.

The Treasury Department has gone a step further. Under the DATA Act, the existing USASpending.gov spending data portal must publish a more complete and detailed spending data set after the May 2017 deadline. In 2015, Treasury began testing new sorting and visualization tools on an experimental website, OpenBeta.USASpending.gov, collecting feedback from the public so that the most useful tools can be applied as soon as the new data set is ready.

"Open data requires both availability and accessibility to democratize data. Governments shouldn't expect much from simply putting spreadsheets on a website - the public needs visualization to create real, usable data."

– Seth Metcalf, Deputy Treasurer at the Ohio State Treasurer's Office

Alongside the growth in open data publication, to the extent allowable by law, agencies are working to share with each other high priority data sets that are closely held to protect individual privacy, industry trade secrets, or national security. For non-public data sets, standardization and better internal sharing can strengthen the ability of agencies to manage programs, conduct investigations, and improve efficiency and effectiveness.

Many of our interviewees were quick to point out that, while data publication has greatly increased, the standardization of federal data has made significantly slower progress (see Perspective: Data Without Standards on page 11). While the DATA Act requires the standardization of all federal spending data, most federal open data published on other platforms lacks such government-wide standardization. As previously discussed, consumers of data cannot make full use of, or comparisons between, data sets that have not been standardized. However, this does not mean published data that is not standardized is without value. In fact, such data can be used in the process of standardization and also to find and correct inaccuracies over time.

Finally, several of our interviewees pointed to a disconnect between government data sets that are released and the data that is useable or needed by data consumers. Agency leadership often decides what data sets to prioritize and release based on internal factors and without adequate input from the private sector, researchers, investors, or even stakeholders within government. In other cases, the data that would be most valuable to users is not released due to technological or systemic difficulties, privacy or security concerns, or the lack of resources required for release.

We asked our interviewees about the importance of data availability and publication in their fields. Here are some of their responses:

BRANDON PUSTEJOVSKY

Chief Data Officer at USAID

Standardization is extremely important, but we are still at an intermediary step -- namely, ensuring the broad availability and machine-readability of our data on a routine basis. Standardization will become more of a priority as we get these initial steps in place. But when moving forward, we can't allow the perfect to be the enemy of the good.

CHRISTINA HO

Deputy Assistant Secretary of Accounting Policy and Financial Transparency at the Department of the Treasury

Data quality is a journey. As more data is published, we will become aware of more problems, but that process will make it possible to improve.
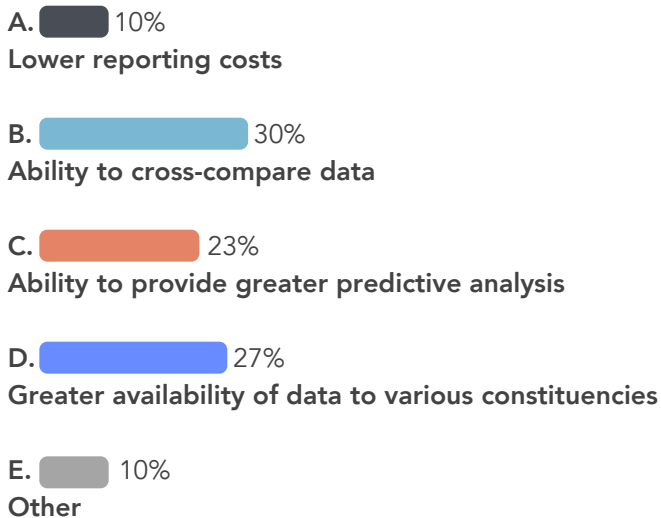
BRYCE PIPPERT

Principal at Booz Allen Hamilton

Open data leaders need to shift focus from just publishing to focusing on releasing granular, real-time value generating data sets.

THE FUTURE OF OPEN DATA

Just as we could never have imagined the Internet of today when the enabling technology was first designed, we cannot clearly picture how our society will be transformed. Our interviewees discussed a diverse set of possible futures, often linked to their fields. The one factor that most agreed upon was timing; we are only at the beginning of the open data revolution, and years will pass before we see the full extent of the society-changing benefits that open data advocates promise.

What do you think will be the most important outcome of more open data in your field?

A. 10%
Lower reporting costs

B. 30%
Ability to cross-compare data

C. 23%
Ability to provide greater predictive analysis

D. 27%
Greater availability of data to various constituencies

E. 10%
Other

The vast majority of our *Data Transparency 2016* speakers said standardization and/or publication of government data relating to their fields has improved to at least some extent over the last few years, some characterizing open data as a "radical transformation" or a "monumental change." However, a vocal minority disagreed, treating open data as a "buzzword" or a transient industry trend.

Some interviewees discussed relatively modest steps forward for agencies in the short term. Understanding that legacy systems can make the process of data extraction and publication resource intensive, they highlighted the need to build in open data processes within new programs and systems natively, when they are first created. These interviewees suggested what we now refer to as "open data" will, in the future, simply be part of good government management, rather than a separate concept.

Other interviewees described a longer-term vision that was more expansive. For example, a few experts associated open data with the **Internet of things** (the connections between the Internet and tech-enabled objects in the physical world, like alarm systems, cameras, cars, *etc*.). Other interviewees described a future that is more pragmatic, one in which we are able to manipulate and make comparisons across vastly different types of data in order to achieve remarkable insights. In such a future, we may be able to track federal dollars from Congressional authorization through the federal agencies all the way to specific programs and payments, or determine the number of people receiving assistance from multiple federal programs and assess whether those people are being well-served. We might be able to analyze data from previous natural disasters to understand exactly the resources needed to serve the affected population of a new one, or make accurate predictions for agriculture subsidy programs based on global weather and economics.

There was an interesting contrast between interviewees who were broadly optimistic about the promise open data and those who viewed the concept with some cynicism. It is possible much of this contrast exists because the concept of open data is not fully understood by everyone who uses the term. Agency staff and the public may hear leaders frequently use the term "open data" without context or with nearly-utopian promises.

The difference between some interviewees' ambitious visions and others' skepticism might also result from the lengthy delays faced by the government's most valuable open data projects. The DATA Act was first introduced in Congress by Rep. Darrell Issa and Sen. Mark Warner in June 2011; a single open data set covering all federal spending will not exist until May 2017, nearly six years later. The Securities and Exchange Commission decided to begin requiring public companies to report financial statements as open data in 2009; seven years later, nearly all companies still prepare two separate versions of each statement – an open data version and a document version. In general, the more valuable a particular compilation of information would be if transformed into open data, the greater are the programmatic and technological barriers to doing so.

Some of our interviewees discussed how we can expect government agencies to use open data in the short-term future:

JOEL GURIN

President of the Center for Open Data Enterprise

It's realistic to ask that government agencies 1) develop accurate open data inventories, 2) identify high priority data, and 3) make high priority data available in accurate, accessible, interoperable formats.

BETH BLAUER

Founder at the Johns Hopkins University Center for Government Excellence

Government agencies are themselves the prime users of government data. Data should be recognized as a strategic asset, and therefore a data strategy connected to real goals and outcomes is critical.

While open data holds a great deal of promise to make government more efficient and programs more effective, there are very real challenges, especially data standardization, which must be met in order to realize its full potential.

Some interviewees offered specific suggestions of the fields in which open data can deliver specific benefits:

AUDITING AND FRAUD DETECTION

The increased availability of standardized financial data is already significantly augmenting the work of government investigators and auditors to identify fraud and waste. The same techniques and tools are available for both internal and external investigations, whether conducted by inspectors general hunting public-sector fraud or agencies like the Securities and Exchange Commission and Department of Justice investigating private-sector abuses.

Even when financial data may not be open to the public for reasons of privacy or security, government investigators may access it, make comparisons with other data sets, and more quickly and easily spot inconsistencies and errors. But open data is even more valuable because oversight can be crowd-sourced.

As the government replaces document-based records with standardized data filings, it will enable grantees, contractors, and regulated entities to report financial information automatically. Ultimately, automation will reduce compliance and transaction costs, allowing for more real-time monitoring of finances and detection of inconsistencies.

"Bernie Madoff was able to fool financial regulators because they are still using 1930s pen and paper technology to handle today's digital challenges. If the regulators adopted standardized data fields and formats across the information they are already collecting under existing financial laws, data analytics could catch fraudsters and inform better decisions."

– Representative Randy Hultgren (R-IL)

MARKET DEVELOPMENT AND COMPETITION

Several of our interviewees pointed out the importance of open data for investment. For example, investors might rely on open data regarding municipal debt and infrastructure to plan where to invest; they might back projects or companies that use open data in innovative ways; or they might rely on more accurate, available, and comparable financial data about potential investment choices.

Similarly, a number of our interviewees discussed how open data can flatten the playing field in many sectors, allowing for new market entrants and for smaller businesses to compete with larger ones. Among government contractors, for example, information about government spending and contracts is a key strategic resource. Companies spend significant time and resources acquiring such information. Once spending data sets are fully open, contractors will be able to analyze them more cheaply, understand trends in pricing and other areas, and make their bids more competitive, all while spending fewer resources on information gathering.

> The challenges holding back the release of open data are not technological. Unlocking the value of open data requires changes in mindset, culture, and priorities. We have some good foundational policies, but is government really ready for the more difficult changes that will lead to value creation for society?
>
> – Bryce Pippert, Principal at Booz Allen Hamilton

> "Rather than limiting data to a privileged few, publication opens data to all market participants."
>
> – Mike Starr, Senior Vice President of Government and Regulatory Affairs at Workiva

LEGISLATION AND POLICYMAKING

The publication and standardization of legislative and regulatory information is at an early stage compared to many other open data domains. Most states publish laws and bills in closed formats, and the field has a number of proprietary services that use limited data-scraping or human analysts to comb through bills and regulatory rules as they are created. But what could be more central to the functioning of our democracy than public access to legislative materials, expressed as open data? Open data would allow lawmakers and advocates to quickly and easily compare laws across states, enable businesses to quickly understand how regulations might apply to their work, and give citizens instant insights into their representatives' work – with opportunities to influence it.

CONCLUSION

What is needed to make all these promising visions our new reality? With reforms such as the Executive Order 13642 and the passage and implementation of the DATA Act, policymakers' understanding and government's very culture regarding information has already begun to shift. The culture must shift further.

Agencies are considering how to make their data more available and usable much earlier in the course of program development than they used to. To this extent, open data is already a vital part of the conversation.

At the same time, however, progress on standardization has been limited and incremental. The DATA Act and the SEC reporting requirements demonstrate a need for clear responsibility to collaboratively create, and strongly enforce, data standards. But for many valuable information resources beyond spending and corporate finance, there is no clear resonsibility to standardize. Government can and should fix that by appointing more standard-setters.

Even as the standardization problem is solved, several of our interviewees reminded us that data is just a resource, not an end in itself. Data will only tell us the answers to questions we think to ask. Our large, diverse society's differing viewpoints will always filter the story that open data can tell.

As we move into the era of open data, we are reminded that all government information ultimately belongs to the People, and the caretakers of open data have a duty to use it wisely and well.

It is so easy to believe data will solve all our problems. What is the point of data if citizens don't trust the government? Citizen engagement is critical - we need to ask for what and for whom are we collecting data.

– Sonal Shah, Executive Director at the Beeck Center for Social Impact & Innovation

APPENDIX

We thank all the government officials and policymakers, industry leaders, experts, and advocates who agreed to interview with us or provide their perspectives for this project.

Steve Adler	Seamus Kraft
Perryn Ashmore	Karen Lay-Brew
Gary Bass	Mark Meadows
Beth Blauer	Kevin Merritt
Robin Carnahan	Seth Metcalf
Francesco Ciriaci	Jason Miller
Amy Deora	Dan Morgan
Jonathan Elliot	Seth Moulton
Mark Flannery	Phil Moyer
John Gentile	Bryce Pippert
Terry Grafenstine	Brandon Pustejovsky
Joel Gurin	Matt Reed
Frans Heitbrink	Tammy Ripperda
Christina Ho	Adam Roth
Hudson Hollister	Sonal Shah
Alex Howard	Robert Shea
Adam Hughes	Mike Starr
Bill Hughes	Rob Surber
Randy Hultgren	Kelly Tshibaka
Waldo Jaquith	Reed Waller
Craig Jennings	Sherry Weir
Marc Joffee	Dave Williams
Taavi Kotka	Mike Willis