HYBRID DATA
and the Promise of a Modern Digital Government

Richard Beutel, Esq.
Principal, Cyrrus Analytics LLC
About the Author

Richard Beutel is the founder and leader of Cyrrus Analytics.

A nationally recognized expert in IT acquisition management and cloud policy, Richard Beutel has 25 years of private sector experience and was the founder and co-managing director of the Dell Government Relations Team.

Most recently, Beutel was appointed Director of the prestigious Procurement Roundtable. He is also on the Executive Committee of ACT-IAC, working to implement reforms in IT acquisitions and to accelerate the cloud first mandate.

Richard Beutel is a leader in advancing the original Cloud First Policy and pushing to modernize IT across the federal government. He authored the original text of the Modernizing Government Technology (MGT) Act working directly with Senate staffers to push forward this essential proposal to modernize IT across the federal government.

Beutel is the original author and legislative manager for the Federal IT Acquisition Reform Act (FITARA). FITARA was signed into law in December 2014, and provides Chief Information Officers (CIO’s) with unprecedented budgetary and legal authority over the operation and deployment of critical Information Technology systems across the federal government.

Previously, Beutel served for over 10 years on Capitol Hill. Most recently, he was the lead acquisition and procurement policy counsel for Chairman Darrell Issa of the House Oversight and Government Reform Committee. He has bicameral Congressional experience, previously serving as lead oversight and acquisition policy counsel for Senator Susan Collins, the ranking member of the Senate Homeland Security and Government Affairs Committee.
I. Introduction

The digital revolution, ubiquity of the internet, and rise of Big Data have given government an unprecedented capability to produce, collect, utilize, and disseminate a vast array of information and data. These trends have ushered in a new era of data-powered government innovation and citizen services based on the undeniable value in making government data widely available – to citizens, activists, companies, academics, and entrepreneurs. This is often referred to as the “open government” era, which thrives on government transparency, public accountability, and citizen-centered services.

Consequently, the last 20 years have seen a transformation of public policies – legislative, regulatory, and administrative – grounded in the philosophy that access to and dissemination of government data is a public right and that any constraints on access hinder transparency and accountability. While there is broad recognition of the need to maximize access to government data, the types of government data are increasingly diverse and complex. For instance, there are many cases where the government collects or licenses private sector data, often combining this data with other data produced by the government. These datasets are often referred to as “hybrid data” or “privately curated data” – data licensed to or collected by the government that comprises both public and private sources. Access to and use of hybrid data is increasingly critical for government to transform data into actionable information.

Examples of curated, or hybrid, datasets include the integration of traffic-app data with US Department of Transportation information, the incorporation of private geographic mapping software into local government flood tracking, the federal award infrastructure’s use of the Dun & Bradstreet D-U-N-S® Number to administer and oversee a $1.2 trillion federal grant market, and peer-reviewed scientific and technical literature that is based on government-funded academic research but published in the private sector. Subjecting this full range of information to unfettered “openness” requirements risks the availability and quality of these valuable data-driven resources. Such requirements will ultimately harm the public interest when the inevitable “tragedy of the commons” scenario compromises the quality of the dataset, as private-sector actors begin avoiding these government partnerships for fear losing control of their data.

Unfortunately, some current open data policies invite unintended consequences – specifically, well-intentioned but overly broad open data mandates that nullify intellectual property rights by extending to data produced in the private sector and collected by, or licensed to, the government. In these cases, the pursuit of maximum data-driven transparency often conflicts with other important public-interest goals, such as rewarding data-driven innovation, safeguarding individual privacy, protecting intellectual property, encouraging private-sector innovation, and promoting the government’s access to data-driven tools that enable smarter decision-making.

Therefore, policies and requirements for openness of government data must contend with these unique challenges and take care to avoid unintended consequences. To be sure, resolving these tensions is not easy, as it requires the nuanced balancing of competing public interests (e.g., effective and accessible government versus open government), but it is possible – and urgent. This paper offers a framework for addressing the hybrid data issue in a way that maximizes the transparency and accountability goals of the open government data movement, while avoiding unintended negative consequences. This paper will:

- Review the history of federal open data policies
- Assess the challenges posed by hybrid data
- Suggest principles for better open data policies and next steps for policymakers

There are many cases where the government collects or licenses private sector data, often combining this data with other data produced by the government. These datasets are often referred to as “hybrid data” or “privately curated data” – data licensed to or collected by the government that comprises both public and private sources. Access to and use of hybrid data is increasingly critical for government to transform data into actionable information.
II. History of Federal Open Data Policies

The digitization and dissemination of government information via the internet began in the early 1990s. In 1994, the Government Printing Office (GPO) launched the GPO Access website (now FDsys), which provides the public with electronic copies of the Congressional Record, texts of bills before Congress, and provisions of the United States Code.¹

That same year, the Clinton administration issued the Circular A-130 memorandum, which championed the importance of providing government information to the public by establishing the first federal Information Management Policy.² In 1995, the House of Representatives created THOMAS.gov – now Congress.gov – a website that provides the public with comprehensive, up-to-date information about pending legislation before Congress.³

The above examples are only a subset of the larger federal shift toward electronic openness, and while these early actions represent landmark moments in technological and political history, the modern open government movement found renewed and more expansive success following the election of President Barack Obama. The Obama administration’s commitment to core open government principles led to enactment of major policy changes through both executive and legislative actions:

a. Executive Actions

On his first full day in office, President Obama issued the Open Government Directive, with the goal of ushering in an era of open government that would “bridge the gap between the American people and their government” by “taking specific actions to implement the principles of transparency, participation, and collaboration.”⁴ The subsequent rollout in late 2009 of Data.gov represents one of the most notable actions undertaken in support of the Open Government Directive. Data.gov aggregates and provides public access to feeds of government data. At the time of publication, the current total of datasets available to the public is 226,298, but this number rapidly grows each day.⁵

In his 2010 address to the United Nations General Assembly, President Obama described open governments as the “strongest foundation for human progress” and called on the member nations to return next year with “specific commitments to promote transparency; to fight corruption; to energize civic engagement; [and] to leverage new technologies so that we strengthen the foundations of freedom in our own countries[] while living up to the ideals that can light the world.”⁶ One year later, the United States joined seven countries to launch the Open Government Partnership (OGP) – a multilateral initiative to support national efforts to promote transparency, fight corruption, strengthen accountability, and empower citizens.⁷ Since its creation, over 70 OGP countries and 15 subnational governments have made over 2,500 commitments to promote transparency and make their governments more open.⁸

The Obama Administration brought further clarity to the domestic open government movement by characterizing open data as a critical national asset in the M-13-13 memorandum establishing the Open Data Policy pursuant to the 2013 Executive Order (EO) “Making Open and Machine Readable the New Default for Government Information.”⁹ The Open Data Policy set forth a list of governing principles for the government’s use and dissemination of open data – defined as “publicly available data structured in a way that enables the data to be fully discoverable and usable by end users.”¹⁰ In conformity with this policy, open data is to be public, accessible, described, reusable, complete, timely, and managed post-release.¹¹

---

1 See https://www.govinfo.gov/about/history (Accessed March 8, 2018).
3 See https://www.govinfo.gov/about/history (Accessed March 8, 2018).
7 See https://www.opengovpartnership.org/about/about-ogp (Accessed March 8, 2018).
8 Ibid.
9 President Barack Obama, Memorandum to the Heads of Executive Departments, M-13-13, May 9, 2013, “Open Data Policy – Managing Information as an Asset.”
10 Ibid.
11 Ibid. This last requirement – the management of data post-release – is an important reason why curated data must be protected in many instances, otherwise there is no economic or financial incentive to manage the data post-release.
II. History of Federal Open Data Policies (cont.)

Executive Actions (cont.)
Currently, President Donald Trump and his administration have publicly affirmed their support of the open government principles espoused by the previous administration. In addition to hosting an Office of Management and Budget (OMB) roundtable on open data, the Trump administration also committed to the OGP by introducing a draft framework for the Fourth U.S. National Action Plan for Open Government. These efforts have been complemented by the Trump administration’s information technology modernization goals, including the publication of a Federal IT Modernization Report and a General Services Administration (GSA) workshop on “Emerging Technology and Open Data for a More Open Government.” Notably, the GSA framed this workshop on the understanding that “[o]pen data and emerging technologies – including artificial intelligence and distributed ledgers, such as blockchain – hold vast potential to transform public services held back by bureaucracy and outdated IT systems.”

b. Legislative Actions
Several legislative actions pursued during the past decade have extended these executive commitments to open government.

In 2014, Chairman Darrel Issa and Ranking Member Elijah Cummings of the House Oversight and Government Reform Committee championed the landmark Digital Accountability and Transparency Act (“DATA Act”). The DATA Act requires federal agencies to report financial spending data in machine-readable formats. President Obama signed the DATA Act in the spring of 2014, and federal agencies first issued financial spending data in machine-readable form in the early summer of 2014. The legislation requires that every federal agency must begin reporting open spending data by May 8, 2017; not all agencies met the deadline, but efforts to comply with DATA Act requirements continue apace.

The evolution of the open government movement made another critical advance in 2017 with the issuance of a report by the Commission on Evidence-Based Policymaking. Chartered by Public Law 114-140, the Commission was directed by Congress to examine and issue recommendations “to study how the data that government already collects can be used to improve government programs and policies.” In this report, the government acknowledged for the first time that a focus on public open data is necessary but not sufficient for good digital governance. Instead, it is the synthesis of all types of data into actionable policy-driving evidence that must now become the focus of the open government movement.

…the United States joined seven countries to launch the Open Government Partnership (OGP) – a multilateral initiative to support national efforts to promote transparency, fight corruption, strengthen accountability, and empower citizens.

70 OGP countries | 15 subnational governments | over 2,500 commitments to promote transparency

---

16 Ibid.
18 Ibid.
II. History of Federal Open Data Policies (cont.)

Legislative Actions (cont.)
The Commission expressly called for legislative changes to improve data quality, data use, and data deployment as critical tools for better policymaking.\textsuperscript{21} Congress adopted the work of the Commission in proposed legislation: Foundations for Evidence-Based Policymaking Act of 2017, H.R. 4174.\textsuperscript{22} This bill also incorporates the text of H.R. 1770, the OPEN Government Data Act, which requires open government data assets made available by federal agencies to be published as machine-readable data. Introduced by Speaker Paul Ryan on October 31, 2017, H.R. 4174 cleared the House Oversight and Government Reform Committee on November 15, 2017, and was passed by the House under suspension of the rules on November 16, 2017. The bill currently awaits Senate action in the Homeland Security and Government Affairs Committee.

c. Relevant Technology Advances
The focus upon specific, standards-based data formats is rapidly becoming less important as the result of an entire generation of “middleware” or “data layer” technologies that are capable of absorbing data in any (or in no) specific protocol or format and assimilating that data directly into robust data analytics products. Agencies are aggressively implementing so-called “data layers” to automate and improve existing business systems while running on top of legacy datasets. One such example is HHS, where the agency is implementing middleware to create a data layer running on “top” of existing legacy platforms and datasets.\textsuperscript{23} Automation and blockchain are key to ensuring availability and use of data, regardless of its individual format (or even if it lacks a specific format because it is completely unstructured). In this fast-paced technological environment, writing a technical data format standard or other “tech mandate” into law makes no sense. Such requirements are very expensive to implement and at risk of becoming moot (or even counterproductive) in the blink of an eye.

\textsuperscript{21} Ibid, 19.
\textsuperscript{22} Foundations for Evidence-Based Policymaking Act of 2017, H.R. 4174.
\textsuperscript{23} Blockchain, automation and Learning to love your legacy systems; Troy Schneider, FCW (Apr 13, 2018).
III. Challenges for Open Data Policies

A key pillar of the open government data movement is the requirement that the government must provide access to its data exclusively in open, non-proprietary data formats that are unencumbered by restrictions on their public release. This is a laudable goal and unobjectionable when applied to purely government data, but it fails to recognize the realities of the hybrid data environment.

For many years, governments have collected, utilized, and relied on a range of non-public data from sources other than government itself, whether for reasons of cost savings, privacy, efficiency, or added value. In some cases, making the best decisions and achieving better governance requires the government to work with a mixture of public and private data – and these sometimes come with appropriate restrictions on their public release.

Therefore, open government data policies must not be founded on the underlying belief that transparency trumps all other considerations. While maximal transparency is indeed essential for effective public oversight of government, good government has other essential ingredients, such as improving government performance and citizen services, maximizing taxpayer value, and ensuring policymakers and bureaucrats are equipped with accurate data that drives innovation and better decision-making. Transparency is an important byproduct of – but not the reason for – government utilization of data.

Collecting, verifying, analyzing, and publishing accurate datasets is a resource-intensive activity that generates valuable assets and solutions which governments need. This effort demands time and money and manages several competing interests, including individual privacy, national security, and intellectual property. Entities – both private and public – who engage in this economic activity prefer not to have the fruits of their investment publicly released in a way that would undermine their value. Yet that is what some open government advocates appear to be demanding as a blanket rule – a rule that, if followed to its logical conclusion, could discourage or eliminate public-private data collaborations that result in enormous benefit for the government and taxpayers alike.24

a. Government Reliance on Proprietary Restrictions

The government acknowledges the value of certain data restrictions necessary to capture the economic value of data it curates at taxpayer expense. For over 50 years, the National Institute of Science and Technology (NIST) has developed and distributed Standard Reference Data in chemistry; engineering; fluids and condensed phases; material sciences; mathematical and computer sciences; and physics.25 These carefully curated datasets are proprietary to NIST under P.L. 90-396, the Standard Reference Data Act,26 and as such the public does not have blanket, free access to this government data. These datasets are sold to the scientific community under copyrights held by the US Secretary of Commerce on behalf of the United States. While some of the Standard Reference datasets are provided to the public for free, many are not. Royalties received for these curated datasets are used to defray the costs of publication and curation and to maintain data quality and availability.

b. Opportunities for Public/Private Collaboration

If private companies forgo data-collaboration partnerships with the government because of blanket open government data requirements, the government will be forced to create, develop, and continuously invest time and money into building its own datasets, instead of leveraging private datasets that are better, more accurate, and more cost effective from private companies that have been building and improving these datasets over decades.

---

24 Perhaps the clearest example of establishing an environment where public–private partnerships may flourish is the Defense Departments's establishment innovation cells, such as the group called DiUX. Innovation cells are authorized to use a special procurement technique known as “Other Transaction Authority (OTA),” OTA allows industry partners to keep and protect their intellectual property rights (IPR) in ways that traditional government contracting forecloses. Industry is therefore more willing to partner with government and share their innovations. Without OTA and reservation of industry IPR, it is certain that many of these innovators would refuse to partner with government.


26 Standard Reference Data Act, Public Law 90-396.
III. Challenges for Open Data Policies (cont.)

Opportunities for Public/Private Collaboration (cont.)

The government is often not best suited to produce quality curated data. For example, USAspending.gov – the online portal that provides the public with spending data for federal agencies – has experienced various implementation difficulties since its debut in 2007.27 In 2011, the OMB contemplated shutting the portal down because the government could not afford the $4 million in operating costs.28 A 2014 Government Accountability Office (GAO) report discovered that the portal failed to report on $619 billion dollars’ worth of federal grants and awards.29 As a result of these continued financial and functionality issues, Congress included a mandate in the 2014 DATA Act for the Treasury Department to upgrade the portal.30 Following these improvement efforts, Treasury unveiled a costly beta upgrade in 2017.31 According to the OMB, the total investment costs of improving and operating USAspending.gov in fiscal year 2018 will amount to $14.3 million.32

USAspending.gov is not the only example of an excessive use of taxpayers’ dollars for an insufficient data-curation service. A December 2017 report by the GAO highlighted the shortcomings of the Transformed Medicaid Statistical Information System,33 The Centers for Medicare & Medicaid Services launched this initiative in 2011 to improve Medicaid data. Unfortunately, significant functionality complications have stymied improvements. The GAO reported incomplete data, inadequate data standards, and an estimated $36.7 billion in improper payments during fiscal year 2017.34

The government has long understood the challenges involved in supplying itself through internally produced goods and services, which is why OMB Circular No. A-76 sets forth the proper delineation between private and public work. The government recognizes that:

[t]he competitive enterprise system, characterized by individual freedom and initiative, is the primary source of national economic strength. In recognition of this principle, it has been and continues to be the general policy of the Government to rely on commercial sources to supply the products and services the Government needs.35

In addition to efficiency issues, a rule mandating government datasets be made available to the public without any restrictions could also cause unintended harm to privacy and national security. In instances where government and private datasets are blended, the government should work with private companies to find the proper balance through the development of the most effective programs with the most accurate data.

c. Existing Use Cases for Private–Public Partnerships

The US Department of Transportation (DOT) recently announced a partnership with the navigation app Waze on a pilot project aimed at decreasing fatalities on US highways.36 The DOT intends to leverage Waze data on traffic hazards and conditions to improve DOT’s own traffic crash data. As the DOT explained in its announcement, “These pilot projects represent a new approach to data analysis that will seek to augment traditional data sources with new data that can be collected and analyzed much more quickly. This approach will create new multidimensional models of the transportation system.”37

34 Ibid.
37 Ibid.
III. Challenges for Open Data Policies (cont.)

Existing Use Cases for Private–Public Partnerships (cont.)

It is unlikely that this project will move forward if the government mandates that Waze deed its company datasets to the public domain. In fact, Waze has restricted the distribution of its proprietary data in past collaborations with state transportation departments, and it is expected that there will be similar restrictions on the data that Waze shares with the DOT.38 Similarly, scientists and officials in Virginia Beach recently initiated a private–public partnership with the creation of StormSense – a web platform that incorporates the software company Esri’s geographic mapping software with data from the National Institute of Science and Technology (NIST) and the U.S. Geological Survey.39 The platform is designed to provide local officials and community members with faster and more accurate information about coastal flooding and sea-level rise. Esri’s platform has helped create a tool to provide more accurate information than that provided in the past by the National Oceanic and Atmospheric Administration.40 The private company Esri is an open data champion that brands itself as “an open platform for innovation” and provides its users with the option of sharing and working with open data, but Esri does not mandate that all datasets created with its platform be open nor does the company release all of its core technologies as open specifications.41 A blanket transparency rule may prevent government from leveraging datasets from a company aligned with most open data advocates.

The federal government’s use of Dun & Bradstreet’s Data Universal Numbering System (D-U-N-S) is an important example of the value of hybrid data.42 A Dun & Bradstreet D-U-N-S® Number is a unique individual identifier for any entity that wishes to participate in the government market and is also a critical data asset for the federal government. Like the vehicle identification number (VIN) system, a single, unique data identifier provides significant value for the oversight and management of public contracting. Any government contractor who wishes to participate in the government market can receive this identifier free from Dun & Bradstreet by virtue of Dun & Bradstreet’s GSA contract. The number is then used to identify and track government awards in the form of contracts, grants, and loans, thus streamlining oversight and contract management. While the D-U-N-S Number is a means to uniquely identify award recipients, Dun & Bradstreet’s core mission is to collect and curate data on unique entities, ensure the data’s accuracy and validity, and, ultimately, derive insights into these entities and the relationships between them. This effort essentially provides the government with an outsourced data standards model built upon global standards and laws. Dun & Bradstreet invests more than $1 million every business day toward the maintenance and enhancement of the business information database, including the D-U-N-S Number.43 This investment helps to link information through the D-U-N-S Number about suppliers, customers, and trading partners, which in turn provides the government with comprehensive insight into the risks and opportunities associated with an awarded entity. It is this curation that significantly enhances the utility and value of the Dun & Bradstreet data to government users and has funded multiple innovative offshoots that have dramatically improved government oversight of and public transparency about its contracting practices.44


40 Ibid.


IV. Principles and Next Steps for Better Open Data Policies

A more balanced open data policy should acknowledge and encourage government utilization of the most accurate and cost-effective data available, whether public, private, or a combination of the two. Technology mandates to release all datasets to the public domain fail to serve the broader public interest.

The Obama-era open government policies created a valuable framework for the maximization of purely government data, but these need to be applied carefully and updated to reflect the importance of hybrid data and the growing trends in technology and government data use. In its 2013 Open Data Policy, the Obama administration expressed a *preference* for data formats that are “non-proprietary, publicly available, and [with] no restrictions ... placed upon their use.”

While a preference for openness is appropriate, effective policy should strive to avoid the unintended consequences of moving from preference to technology mandate when applied to hybrid data. Mandates that make public–private initiatives economically unfeasible serve nobody’s interests. Policymakers should caution against statutory restrictions that limit new and innovative technological choices simply because they happen to be proprietary. Instead, a better aim for legislative and regulatory oversight would be encouraging the market and public entities to develop the most efficient and effective data solutions possible while maintaining appropriate openness.

So how do we achieve this balance? The first concrete step toward answering this question should take form in a GAO study of executive agencies’ use of public, private, and hybrid datasets. This study should specifically evaluate current data quality, the tools needed to improve data quality (including curation), and potential improvements to the collection and reporting of government-award data. Fortunately, Congress has already provided the foundation to help guide such an initiative. A GAO study on data quality should build on the work and recommendations of OMB Circular A-76, the Commission on Evidence-Based Policymaking, and the Committee on Oversight and Government Reform’s report on the Foundations for Evidence-Based Policymaking Act of 2017 (H. Rept. 115-411) to ensure the proper roles of government- and private-sector innovations. The Committee on Oversight and Government Reform’s report language on the OPEN Government Data Act is particularly instructive for what propositions a GAO study should test:

Open formats and open licenses are necessary components of a default of openness because they remove barriers to accessing and using the data. The presumptions expand upon, but do not alter existing openness requirements related to the treatment of any work of the United States Government under section 105 of title 17 or any other rights regimes. The default is only that—a default. There are instances where it could be inappropriate for the government to impose open license requirements, such as for data that the government uses and maintains but does not own. For example, an agency might contract with a commercial data provider to obtain data that, if the agency attempted to collect on its own, the agency would need to spend significant time and resources verifying. This bill is not intended to prevent agencies from contracting with commercial data providers to obtain data under restricted terms, when such contract is in the public interest and is the most cost-effective way to meet the Federal Government’s needs.

Studying these propositions will provide the necessary tools to develop a balanced open data policy that achieves the widest possible ease of use and transparency for all forms of data.

---


V. Conclusion

The digital revolution is an opportunity to achieve maximum public transparency and illuminate with sunshine previously opaque government programs and policies.

Legislation that translates the preference for openness into laws and regulations is a positive step, but policy makers should ensure these laws and regulations do not enact a technology mandate with unintended and harmful consequences. The value of data is dramatically enriched by the quality of the data being provided. A tech mandate that effectively shuns the use of curated, or hybrid, data will usher in an era of data that is less reliable, less useful, less desirable, and more expensive. We cannot deprive government of the information and tools it needs to best serve and protect the public. It is time for the government, nonprofit, and private sectors to come together and formulate a balanced approach to the special case of “hybrid data.”