December 14, 2016

The Commission on Evidence-Based Policymaking
Docket ID USBC-2016-000

Re: Comments to the Commission on Evidence-Based Policymaking

To Whom It May Concern:

On behalf of the Center for Data Innovation (datainnovation.org), we are pleased to submit these comments in response to a request for comments from the Commission on Evidence-Based Policymaking, a bipartisan commission created by Congress to examine how to increase the availability and use of government data to build evidence and inform program design while protecting the confidentiality of this data.[1]

The Center for Data Innovation is the leading think tank studying the intersection of data, technology, and public policy. With staff in Washington, DC, and Brussels, the Center formulates and promotes pragmatic public policies designed to maximize the benefits of data-driven innovation in the public and private sectors. It educates policymakers and the public about the opportunities and challenges associated with data, as well as technology trends such as predictive analytics, open data, cloud computing, and the Internet of Things. The Center is a nonprofit, nonpartisan research institute affiliated with the Information Technology and Innovation Foundation.

Over the past several years, the federal government has made substantial progress on making its data freely available to the public in open, usable formats, and this data has proven to be an invaluable resource for informed decision-making. Policymakers should expand on these efforts, as well as address the obstacles that remain to effective data sharing and use.

[1] U.S. Census Bureau, "Request for Comments for the Commission on Evidence-Based Policymaking" September 14, 2016, https://www.regulations.gov/document?D=USBC-2016-0003-0001.

Please find our responses to the relevant questions in the attached document.

Sincerely,

Daniel Castro
Director
Center for Data Innovation
dcastro@datainnovation.org

Joshua New
Policy Analyst
Center for Data Innovation
jnew@datainnovation.org

DATA INFRASTRUCTURE AND ACCESS

BASED ON IDENTIFIED BEST PRACTICES AND EXISTING EXAMPLES, HOW SHOULD EXISTING GOVERNMENT DATA INFRASTRUCTURE BE MODIFIED TO BEST FACILITATE USE OF AND ACCESS TO ADMINISTRATIVE AND SURVEY DATA?

Government data should be open and machine-readable by default, in accordance with official administration policy issued in May 2013.[2] This means that, unless otherwise legally prohibited, government data should use nonproprietary, machine-readable formats, and be licensed to maximize reuse, meaning the data is free for anyone to access, modify, and use for any purpose. Though agencies have made substantial progress towards publishing data as open and machine-readable by default, the requirements to do so are by executive order, rather than an act of Congress, meaning that they could be subject to change under a new presidential administration.[3] Regardless of whether these rules remain on the books, federal agencies should recognize the importance of open data to mission delivery and evidence-based decision-making and continue to treat their data as open by default indefinitely.

Moreover, all federal agencies should produce and publish enterprise data inventories. Enterprise data inventories provide a list of all data assets managed by an agency—both public and non-public. By creating enterprise data inventories, federal agencies, policymakers, and other stakeholders are better able to discover important government data assets.

WHAT DATA-SHARING INFRASTRUCTURE SHOULD BE USED TO FACILITATE DATA MERGING, LINKING, AND ACCESS FOR RESEARCH, EVALUATION, AND ANALYSIS PURPOSES?

The federal open data portal Data.gov is a widely-used and effective tool for making open data from all levels of government easily accessible to members of the public and private sectors alike. Federal agencies should continue to publish their data on Data.gov, ensuring that they publish their data in open and machine-readable formats.

In November 2016, the Department of Commerce partnered with public benefit corporation data.world to use data.world's data science collaboration platform to increase the accessibility,

---

[2] Sylvia M. Burwell et al., "M-13-13," Office of Management and Budget, May 9, 2013, http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf.
[3] Joshua New, "Will Obama be the Last Open Data President?" Center for Data Innovation, November 11, 2014, https://www.datainnovation.org/2014/11/will-obama-be-the-last-open-data-president/.

linkability, and usability of its data.[4] Federal agencies should pursue innovative partnerships such as these that can increase the utility of government data.

WHAT CHALLENGES CURRENTLY EXIST IN LINKING STATE AND LOCAL DATA TO FEDERAL DATA? ARE THERE SUCCESSFUL INSTANCES WHERE THESE CHALLENGES HAVE BEEN ADDRESSED?

Government data published in proprietary formats, data with restrictive licenses, non-machine readable data, data that is not made available free of charge, and data that governments fail to publish are all major challenges for linking data between different levels of government. Unless all levels of government are publishing open data, it can be very difficult for researchers, for example, to know what data is even available to them. And unless data is made available in a nonproprietary and machine-readable format and licensed to maximize reuse, these researchers would be technically and legally unable to link these datasets. In addition, many datasets are not available via application programming interfaces, or APIs. APIs allow developers to access datasets programmatically and expose datasets to new uses, especially on websites or mobile apps.

As noted previously, another important challenge is that most federal agencies have failed to publish enterprise data inventories—a catalog of all of their datasets, both public and nonpublic.[5] Without publicly available enterprise data inventories, members of the public and private sector alike can be unable to locate relevant datasets or determine if a certain dataset even exists.

SHOULD A SINGLE OR MULTIPLE CLEARINGHOUSE(S) FOR ADMINISTRATIVE AND SURVEY DATA BE ESTABLISHED TO IMPROVE EVIDENCE-BASED POLICYMAKING? WHAT BENEFITS OR LIMITATIONS ARE LIKELY TO BE ENCOUNTERED IN EITHER APPROACH?

Data.gov is already a widely-used tool for discovering public administrative and survey data, so there is no need to establish similar tools for public government data. Creating multiple hubs for government data would raise costs for federal agencies, if they must submit data to multiple sites, and for users, if they must search multiple sites. Data.gov could also serve as an index for non-public datasets, but even if they were indexed, these datasets would still only be accessible

[4] Justin Antonipillai and Brett Hurt, "data.world to Bring Valuable Commerce Datasets to 'Social Network for Data People,'" U.S. Department of Commerce, November 8, 2016, http://www.esa.doc.gov/under-secretary-blog/dataworld-bring-valuable-commerce-datasets-social-network-data-people.
[5] Joshua New, "Congress is Stepping Up to Protect Open Data," center for Data Innovation, April 19, 2014, https://www.datainnovation.org/2016/04/congress-is-stepping-up-to-protect-open-data/.

via the security protocols in place to protect the data. Data.gov is not a repository for datasets, and creating a central repository for government data would have limited value because of the costs involved of transferring data from multiple systems. A better approach would be to focus on improving online identification and authentication technologies, so that federal agencies can more easily and securely make non-public datasets available to authorized users. So rather than creating a clearinghouse that is a repository of non-public government data, federal agencies should create a clearinghouse that is an index of available non-public data sources.

WHAT FACTORS OR STRATEGIES SHOULD THE COMMISSION CONSIDER FOR HOW A CLEARINGHOUSE(S) COULD BE SELF-FUNDED? WHAT SUCCESSFUL EXAMPLES EXIST FOR SELF-FINANCING RELATED TO SIMILAR PURPOSES?

In general, the government should refrain from considering its data as an opportunity to generate revenue. Since the government collects data at taxpayers' expense, taxpayers should be able to freely access this information. However, developing the infrastructure to provide large amounts of data to the public can be expensive, so in certain cases, a federal agency may consider charging access fees for certain data, akin to an entrance fee for a national park. However, such fees should solely support the development and upkeep of data infrastructure.

One successful example of a government agency taking advantage of an innovative financing model to build data infrastructure is the National Oceanic and Atmospheric Administration's (NOAA's) Big Data Partnership. Established through a series of Cooperative Research and Development Agreements (CRADAs), the Big Data Partnership is designed to help NOAA publish the large amounts of data it collects—20 terabytes per day, on average—but cannot afford to make publicly available.[6] The Big Data Partnership partnered with major cloud providers, including Amazon Web Services and IBM, to build out NOAA's data infrastructure at no cost to the government, so these partners could access more of NOAA's data to build useful products and services.[7] Importantly, the partner companies do not get prioritized access to this data, as the partnership stipulates all of it must be open.[8] The partners are willing to finance the

[6] U.S. Department of Commerce, "U.S. Secretary of Commerce Penny Pritzker Announces New Collaboration to Unleash the Power of NOAA's Data," News release, April 21, 2015, https://www.commerce.gov/news/press-releases/2015/04/us-secretary-commerce-penny-pritzker-announces-new-collaboration-unleash.

[7] Alexander Kostura and Daniel Castro, Three Types of Public-Private Partnerships That Enable Data innovation," Center for Data Innovation, August 1, 2016, https://www.datainnovation.org/2016/08/three-types-of-public-private-partnerships-that-enable-data-innovation/.

[8] Ibid.

development of this infrastructure because the products they can build with this data, such as improved weather prediction services, will create a substantial enough return on investment.

WHAT SPECIFIC ADMINISTRATIVE OR LEGAL BARRIERS CURRENTLY EXIST FOR ACCESSING SURVEY AND ADMINISTRATIVE DATA?

There are two notable examples of administrative and legal barriers to accessing and using survey and administrative data.

The federal government has multiple agencies collecting the same types of data for economic metrics. The Bureau of Labor Statistics (BLS), Census Bureau, and Bureau of Economic Analysis (BEA) all collect and analyze employment, occupation, income, and other labor market and economic data to produce valuable statistics like productivity growth, state and national GDP, and employment rates that help make policymakers make informed economic policy decisions.[9] This leads to considerable statistical discrepancies between agencies. The issue is that neither of these metrics are necessarily wrong, but the differences in methods used and factors considered by each agency means the same question receives two largely different answers. The root cause of this is that Census can use data that BEA and BLS cannot because some of it is commingled with tax data, which cannot be shared under Title 26 of the Internal Revenue Code.[10] As a result, these statistical agencies cannot rely on an internally consistent dataset.

Another example of a specific barrier is that the Departments of Health and Human Services (HHS), Defense, Education and Justice oversee at least 10 different efforts to collect data about sexual violence—producing widely varying statistics that, on the surface, appear to measure the same thing.[11] As a result, policymakers and the public simply do not have reliable, easy-to-understand data about sexual assault, which can have serious consequences for the effectiveness and accountability of the criminal justice system and hinder efforts to combat sexual assault.[12] The wide variation in reporting can make research about sexual violence unnecessarily difficult,

---

[9] Luke Stewart, "We Have a Sharing Problem," The Information technology and Innovation Foundation, June 13, 2012, http://www.innovationfiles.org/we-have-a-sharing-problem/.
[10] National Research Council et al., *Improving Business Statistics Through Interagency Data Sharing* (National Academies Press, 2006), 15.
[11] Government Accountability Office, *Sexual Violence Data: Actions Needed to Improve Clarity and Address Differences Across Federal Data Collection Efforts* (Washington, DC: 2016), http://www.gao.gov/assets/680/678511.pdf.
[12] Joshua New, "How Common is Sexual Assault in the United States? The Answer Depends On Who You Ask," Center for Data Innovation, September 1, 2016, https://www.datainnovation.org/2016/09/how-common-is-sexual-assault-in-the-united-states-the-answer-depends-on-who-you-ask/.

as comparisons between different groups are all but impossible when methodologies differ so greatly, and the lack of clarity about what these different statistics really mean provides fodder for toxic arguments that dismiss the severity of the problem of sexual assault, such as those that imply that false reporting of rape is commonplace, which is false.[13]

HOW CAN IDENTIFIABLE INFORMATION BE BEST PROTECTED TO ENSURE THE PRIVACY AND CONFIDENTIALITY OF INDIVIDUAL OR BUSINESS DATA IN A CLEARINGHOUSE?

There are a variety of tools and strategies useful for preserving the security and privacy of data. One of the most effective of these methods—de-identification—has been frequently criticized for being unreliable due to the perceived risk that bad actors could easily re-identify a data set.[14] However, while no information security method, including de-identification, is perfect, the risk of reidentification when data is properly de-identified has been greatly exaggerated by some commentators, and it would be a mistake to reverse existing federal policy based on these inaccuracies.[15] The notion that de-identification is an unreliable tool is commonly promulgated by commentators misinterpreting primary literature or failing to recognize that many instances of reidentification were only possible because the data was improperly de-identified in the first place.[16] For example, if a data set is properly de-identified in accordance with the Safe Harbor Standard defined by the Health Insurance Portability and Accountability Act (HIPAA) which requires the removal or modification of 17 specific data elements, only 0.04 percent of individuals are "uniquely identifiable."[17] It is important to note the significant difference between uniquely identifying an individual, which means recognizing a specific, discrete set of characteristics within a data set, and re-identifying an individual, which means gleaning a person's actual identity—his or her name, birth date, and so on—from a data set.

---

[13] Ibid; Ashe Schow, "No, 1 in 5 Women Have Not Been Raped on College Campuses," *Washington Examiner*, August 13, 2014, http://www.washingtonexaminer.com/no-1-in-5-women-have-not-been-raped-on-college-campuses/article/2551980; Carmen Rios, "Rebel Girls: Our 'False Rape' Hysteria is Bullsh*t," Autostraddle, December 31, 2014, https://www.autostraddle.com/rebel-girls-our-false-rape-hysteria-is-bullshit-270299/; "False Reporting," National Sexual Violence Resource Center, 2012, http://www.nsvrc.org/sites/default/files/Publications_NSVRC_Overview_False-Reporting.pdf.

[14] Ann Cavoukian and Daniel Castro, "Big Data and Innovation, Setting the Record Straight: Deidentification Does Work," Information Technology and Innovation Foundation, June 16, 2014, http://www2.itif.org/2014-big-data-deidentification.pdf.

[15] Ibid.

[16] Ibid.

[17] Simon Cohn, "Ltr to Sec'y 2007 Dec 21 Secondary Uses," National Committee on Vital and Health Statistics, December 21, 2007, http://www.ncvhs.hhs.gov/wp-content/uploads/2014/05/071221lt.pdf.

IF A CLEARINGHOUSE WERE CREATED, WHAT TYPES OF RESTRICTIONS SHOULD BE PLACED ON THE USES OF DATA IN THE CLEARINGHOUSE BY "QUALIFIED RESEARCHERS AND INSTITUTIONS?"

In cases where data is inherently sensitive, even after de-identification, other methods should be used to protect data while also making it available to qualified researchers. For example, the Census Bureau operates Federal Statistical Research Data Centers (RDCs), which serve as a secure environment for sharing sensitive data with qualified researchers working on important, pre-approved projects.[18] The Census Bureau collaborates with several statistical agencies to operate the RDCs, such as the Agency for Healthcare Research and Quality and BLS.[19] While the focus of RDCs is statistical agencies, this model could be applied to other government agencies and government datasets with similar effectiveness. In addition, government agencies can require researchers to abide by licensing agreements for non-public data that imposes restrictions on how researchers can use the data.

CONCLUSION

Easy access to useable data plays a crucial role in informed-policymaking, and it is encouraging to see the Commission for Evidence-Based Policymaking working to increase the accessibility and usability of government data for this purpose. Overall, adhering to the principle that government data should be open and machine-readable by default would substantially improve the quantity and quality of data available for evidence-based policymaking.

[18] "Federal Statistical Research Data Centers," U.S. Census Bureau, accessed December 14, 2016, http://www.census.gov/about/adrm/fsrdc/about/available_data.html.
[19] Ibid.