Sharing Securely within Government: Best Practices for Facilitating Interagency Data Science

Kevin H. Wilson
The Lab @ DC
Washington, DC
kevin.wilson@dc.gov

ABSTRACT

The Lab @ DC's mission is to infuse policy making in the Government of the District of Columbia with the best available techniques from social and data science. To do so, The Lab must have access to data sets from many different agencies in the District. However, The Lab realized that there were several barriers to agencies sharing their data. Specifically, there was no common legal or security framework for sharing data. This led the Chief Data Officer to develop a new District Data Policy more clearly defining the types of data the District manages and the responsibilities of the agencies in maintainging that data. Simultaneously, for its own work, The Lab developed a data security policy that all its members must follow when working with data from agencies, as well as a templated data use agreement that allows agencies and The Lab to more quickly agree on how and why data will be shared.

We believe that these problems will be common across governments as they attempt to build and adopt evidence-based policies and data science tools. This short note outlines the principles underlying our solutions, and, as online appendices, makes the text of these policies and agreements available.

KEYWORDS

Security Policies, Data Use Agreements, Data Sharing

ACM Reference Format:
Kevin H. Wilson. 2017. Sharing Securely within Government: Best Practices for Facilitating Interagency Data Science. In *Proceedings of Data Science for Social Good, Chicago, Illinois USA, September 2017 (DSSG'17),* 8 pages.
https://doi.org/10.475/123_4

DSSG'17, September 2017, Chicago, Illinois USA
2017. ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

1 INTRODUCTION

There is an old joke in data science that the field is 80% data wrangling and 20% data science. When it comes to administrative data, the joke really needs a third component: getting data in the first place. For technical, legal, and privacy reasons, it is often nontrivial to gain access to administrative data, which can be both extremely sensitive and extremely difficult to pull out of archaic technology systems.

The Lab @ DC confronted this problem very early in its operations. The Lab, a new group in the Executive Office of the Mayor of the District of Columbia, has a mission directed at bringing the best that social and data science have to offer to the Government of the District of Columbia, but doing so requires access to these sensitive administrative records. For instance, our work helping study the implementation of a nurse triage line[1] in the District requires us to have access to medical records. Our work with the Metropolitan Police Department's recruiting division[2] requires us to have access to the contact information of potential recruits. And our work with the Office of the State Superintendent of Education requires us to have access to student mobility data.

While some of these data can be anonymized, the sensitivity of contact information, medical information, and any information around children means that our partner agencies are rightfully skittish about just handing over reams of data to us. In our early days, we expected this not to be too much of a problem. We expected agencies would have shared their data with numerous people, including other government agencies, and as such would have standard operating procedures in place for sharing data. We found, however, that this was not the case. Instead, much data sharing was being done on an *ad hoc* basis. This is not by itself a problem, but we also found that the *ad hoc* nature of these interactions was causing us quite a bit of trouble as we negotiated with different stakeholders, including program managers, agency directors, and lawyers, from across the District.

The initial solution we have implemented[3] is threefold, and is what we will discuss in the remainder of this paper. The first part is the creation of a District-wide data policy [10]. This was actually not our initiative, but an initiative of the Chief Data Officer in the Office of the Chief Technology Officer. The data policy, signed by the Mayor on April 27, 2017, is meant to standardize terminology around data, including its security and who is responsible for that security. It also establishes a centralized method for gathering metadata about the District's enterprise data sets, and it establishes a framework for proactively publishing public data. We discuss the policy in more detail in Section 2.

The second part of our initial solution is the establishment of a Lab security policy. This policy, developed and implemented with the help of the new office of the Chief Information Security Officer, is a critical component of ensuring agencies that their data will be kept secure. Groups as diverse as the Metropolitan Police Department and the Criminal Code Reform Commission have found its strictures sufficient for data security, and these votes of confidence have helped us secure other data sets. The policy's principles and our implementation are described in Section 3.

The third and final part of our initial solution has been the creation of a template data use agreement. While some agencies have already standardized agreements that govern their own data sharing, we have found that many have not developed standard operating procedures. This is not unexpected since the primary mission of the agencies is not necessarily to share data with outside groups, but to exploit their data for their own operations. On the other hand, data sharing *is* critical to The Lab's operations, and so it made sense for us to create, with our legal team, a standard template on which our sharing can be based. The major principles behind that agreement are described in Section 4.

Finally, we are making each of these documents available on our GitHub page[4] so that other jurisdictions can reuse the work that we have performed. We also welcome feedback on these documents on our GitHub page.

2 THE DATA POLICY

While many cities have recently promulgated open data policies [17] the District of Columbia decided to take a slightly different approach. Working with community and government stakeholders, a decision was made to focus not just on open data, but *all* data. The data policy that resulted was promulgated on April 27, 2017. Chief among its features is giving all agencies in the District a common language to speak about the types of data that are available and the types of personnel necessary to maintain the integrity of those data sets. In particular, it gives a way to speak of how sensitive a data set is, and it anchors that classification to commonly arising legal mandates, such as the Freedom of Information Act [3], Family Educational Rights and Privacy Act [4], and Health Insurance Portability and and Accountability Act [5].[5] Finally, the data policy sets several deadlines for agencies to comply with its terms, most notably cataloging the data sets that they maintain and establishing pipelines for publishing data that should be open to the District's open data portal. In the rest of this section, we discuss some of the highlights of the Data Policy, though the interested reader should peruse the whole document [10].

2.1 The Mission

The Data Policy begins by setting out its mission. Quoting the policy, it requires all District agencies and other bodies under mayoral control to:

[1] Nurse triage lines are an addition to 911 or other emergency systems where if a patient calls with a low-acuity problem, which make up a significant percentage of all emergency calls, they may be redirected to a nurse who can further assist them before sending an emergency vehicle. These lines have shown initial signs of reducing emergency department visits while not decreasing the quality of care in jurisdictions such as Reno [11]. Ongoing updates on our work may be found at https://osf.io/t7nhj/.

[2] We have several ongoing projects with MPD to help improve their recruiting pipeline. Two projects include improving the website of MPD, information about which can be found at https://osf.io/2p7nx/, and evaluating the effectiveness of direct mail on candidate recruitment, information about which can be found at https://osf.io/2p7nx/. Results will be published at these websites as they become available.

[3] This solution is currently being used only by The Lab @ DC and not by the broader Office of the City Administrator or the District Government as a whole. We hope that our work will hone the efficacy of the data policy, the security policy, and the template data use agreement so that they may eventually be deployed more broadly within government.

[4] https://github.com/thelabdc/PAPER-DSSG-2017

[5] The first of these laws, FOIA, requires that certain government records be made available at the request of members of the public. The latter two, FERPA and HIPAA, place restrictions on the use of, respectively, education data and health data to protect the privacy of the individuals whose data is being used.

- Maintain an inventory of its enterprise data sets;
- Classify enterprise data sets by level of sensitivity;
- Regularly publish the inventory, including the classifications, as an open data set; and
- Strategically plan and manage its investment in data.

These four points require the District to centrally understand what data sets it is actually collecting, which will allow the District as a whole to coordinate its investments much more efficiently than it does now. Moreover, it is not uncommon for the data necessary to evaluate a program or policy implemented by one agency to be housed in a different agency. For instance, the District, through its Office of Victim Services and Justice Grants(OVSJG), provides funding to some community-based organizations whose goal is to reduce the truancy rate of kids in the District's public schools [12, 16]. However, the data necessary to evaluate the efficacy of such programs, including attendance and demographic data, is not held by OVSJG, but rather by the various schools in the District. For OVSJG even to plan a study of its own programs, it would be extremely helpful to know exactly what data are available, and especially any quirks they are likely to come across.

The District's data policy also stipulates that, to the fullest extent possible, "enterprise data sets shall be open by default, published online" in a machine readable format without license restrictions or costs to the users, and documented to make the use of the data set relatively easy [10, §I.C]. This emphasizes that open data is a central aim of the District data policy. Specifically, it places a large value on proactively sharing any data that can be publically shared and proactively determining what data cannot be publically shared. While not written into the policy, this has also spurred much work on the part of the Office of the Chief Technology Officer to build standardized data pipelines so that agencies can easily ship their data to the open data portal in a timely fashion.

2.2 Security Levels

There are many, many different laws that limit the use of various government records. Certain criminal justice information, especially expunged and juvenile sentences, cannot be shared publically. HIPAA guards certain health records. FERPA guards certain educational records.

On the other hand, several laws demand that government data be made public. The minutes of meetings of the Council of the District of Columbia must be public, and similarly the meetings of various commissions in the Executive Branch. Moreover, many of the records that the District produces are subject to the Freedom of Information Act, which requires that, upon request, the government provide copies of records pertaining to a particular topic. Though, FOIA also has many exemptions, notably for "deliberative process," which can be quite complicated.

When talking among users of data in the entire District of Columbia government, it can thus be extremely difficult to determine how sensitive the data being discussed is. The data policy attempts to clear this up by setting five levels of data sensitivity, numbered 0 through 4.

Level 0 data is public data, and should be proactively disclosed on the District's open data portal under a Creative Commons CC0 Universal License [2]. Such data include corporate registrations, business licenses, road layouts, trash can locations, anonymized 311 requests, and other non-sensitive data.

Level 1 data is data which is subject to public disclosure under FOIA, but which the District chooses not to release proactively. The prototypical example of such a data set is the voter rolls maintained by the District's Board of Elections. These rolls, which by law are public data, contain all voters' names, addresses, party affiliations, and whether or not they voted in recent elections. Proactively disclosing such data could be dangerous to various vulnerable populations, e.g., survivors of abuse or individuals who are at heightened risk from criminal activity. Thus, the District classifies this data set as Level 1, not proactively released.

Level 2 represents non-sensitive data which is for District government use only. These data sets are often purchased from other entities who require, by license, that the data not be disclosed to the public. For instance, The Lab has a subscription to a journal service and frequently downloads academic articles. This data set of articles, available to anyone in the Office of Performance Management, cannot be published as that would violate the intellectual property rights of the publishers.

Level 3 consists of data which is either highly sensitive or is "lawfully, regulatorily, or contractually restricted from disclosure to other public bodies." That is, this represents data that cannot be freely shared among different agencies of the District government. Such data includes, for example, certain criminal justice information, protected health information, certain education data, and sensitive law enforcement data. Protecting this level of data is what The Lab targets with its security policy and template data use agreements.

Finally, Level 4 consists of data the disclosure of which may cause "major damage or injury, including death, to residents, agency workforce members, clients, partners, stakeholders, or others identified in the information, or otherwise significantly impair the ability of the agency to perform its statutory functions." A typical example of this sort of data is the license plate numbers of the Metropolitan Police Department's unmarked vehicles, which, if disclosed, could turn those vehicles into easy targets. When working with such data in the future, The Lab plans to work on a case-by-case basis with agencies to make sure that our protections are adequate.

2.3 Personnel

Finally, the data policy creates new roles in each agency whose mission is to see that the policy is faithfully executed. In particular, the policy requires the establishment of two titles in each agency: Agency Data Officer and Agency Information Security Officer, and at OCTO, the policy establishes the offices of the Chief Data Officer and Chief Information Security Officer to coordinate the efforts of their agency counterparts.

The data officers, in coordination with the Chief Data Officer, are responsible "for the District's data governance processes, including the collection, creation, maintenance, documentation, dissemination, and archiving of high-quality, highly interoperable datasets." The exact contours of this role will differ by agency, but it is expected that this will include coordinating with OCTO's data team to build pipelines which automatically update a centralized data

warehouse, and will also include simply making sure that the public inventory of data sets is up-to-date.

The information security officers, in coordination with the Chief Information Security Officer, are responsible for "implement[ing], manag[ing], monitor[ing], and report[ing] on cyber security for their assigned agency" and, more generically, coordinating "information security strategy and practices."

3 THE SECURITY POLICY

As we saw in the previous section, data can require vastly different amounts of protection. Some data, like property records, are completely open and you can find them on the District's open data website [13]. Others, like the license plate numbers of the Metropolitan Police Department's unmarked vehicles, are extremely sensitive and must be handled with great caution.

In preparation for receiving and using data requiring all levels of vigilance, The Lab constructed a security policy that addresses many of the concerns that come along with holding sensitive data. Before members of The Lab access a partner agency's data, a signed copy of the security policy is given to the agency.

The policy itself focuses on three prongs: encrypting data as it is transferred, encrypting data when it is not being used, and hardening how we authenticate with protected resources. For the rest of this section, we cover each of these things in detail.

3.1 Encrypting in Transit

No matter how hardened your main systems are, when you share your data with a third party, it is no longer bound by the security of that system. As such, the first principal of The Lab's security policy is that data must be *encrypted in transit*, that is, as it is being transferred from a partner agency to The Lab, it must remain encrypted.

In our experience outside of the District government, we have commonly seen people email personnel data, educational records, and even medical data. But, as usually configured, email isn't secure at all. Your systems administrator has access to the email, it is often permanently archived even when the data itself should be deleted, and it is rarely encrypted.

A typical workaround is for people to transfer data on a flash drive, but this is inconvenient as you must physically go to a computer which has access to the original data, and if the flash drive is not a fresh, encrypted flash drive, it is extremely insecure. If the flash drive is lost, then the data could be anywhere. If the data is not securely erased from the flash drive, it could linger for years and eventually get lost. And it is not at all uncommon for flash drives to be vectors for malicious software to get onto secure networks (for example, see [9]).

A long known solution to this problem is to use SFTP.[6] While this is a complete solution, it is quite complicated to get started. In particular, the agency providing data must download special software to share the data with The Lab. Since many District computers do not allow their users to install new software, this was an unacceptable solution for us.

To solve this problem, The Lab built a small web application that operates over the secure HTTPS protocol and allows Lab members to create secure drop boxes for our partners to upload data. The partner can upload the data by going to a website only accessible on the District's internal network, fill in a short-lived token that The Lab provides, and upload their data. The Lab member who created the secure drop box is then the only person who is able to see that data, but they may provide short-lived access to the data to others.

We call this service the Secure Uploader[7]. We look forward to blogging about its details and how to set it up soon, but for now, feel free to visit its GitHub repository and peruse.

Of course, there are many other services available which will provide similar (and, to be honest, more robust) functionality. The Lab looked into services such as Box[8] and DropBox[9], but ultimately decided to build our own. This decision was made in large part because we did not know whether our partners would balk at storing sensitive data in a third-party service. As we work with more agencies and gauge their willingness to use such services, we plan on revisiting this decision since third parties provide much greater functionality and reliability than anything we can provide internally.

3.2 Encrypting at Rest

The second principle that underlies The Lab's security policy is that even when data is not being used, it should remain encrypted. This principle is known as *encryption at rest*. For The Lab, the data which we are custodians of lives in two major places: our District-issued laptops and on the machines that we rent from Microsoft using their Azure cloud service. In both cases, we make sure the hard drives on which data is being stored is encrypted, though each requires a different approach.

3.2.1 Encrypting our Laptops. The first problem is making sure that our District-issued laptops' hard drives are encrypted. For a long time this has been a built-in feature of Linux and Mac machines (using dm-crypt and FileVault, respectively). On Windows machines, encrypting the hard disk has typically required buying third-party software, but with the advent of Windows 10, Microsoft provides a built-in solution called BitLocker.

While these services are built-in, managing them can be somewhat tricky in an enterprise setting. In particular, in an enterprise setting, it is important that the systems administrators be able to gain access to employees' files. This access can be necessary for compliance and regulatory purposes, as well as for recovering data in the unfortunate event that an employee leaves the enterprise.

Central to solving this problem for The Lab has been working in coordination with the Office of the Chief Technology Officer, especially their IT and security staff. In particular, the data policy's establishment of a Chief Information Security Officer has been a great boon to The Lab's security efforts, as well as those efforts across the District's government.

[6] It is important to note that SFTP and FTP are different. FTP is an older, insecure protocol where anybody monitoring the conversation between the receiver and the sender can reconstruct the files shared. On the other hand, SFTP makes sure to encrypt the files being sent in such a way that they cannot be reconstructed by anybody listening in. If you are using an FTP server to share files currently, we recommend you ask your IT department about implementing SFTP instead.

[7] https://github.com/thelabdc/LAB-SecureUploader
[8] https://box.com
[9] https://dropbox.com

3.2.2 Encrypting our Cloud Machines. In addition to our individual machines, we also utilize cloud computing resources through Microsoft's Azure. We use this offering especially for our shared computing environment which is built on top of JupyterHub. In particular, we use their Government Cloud to maximize our compliance with various federal and local regulations.

In Azure's commercial cloud offering, Microsoft makes encrypting disks extremely easy with their Managed Disks offereing. However, at the time of writing, Managed Disks were unavailable in the Government Cloud.[10] It is still completely possible to encrypt data disks in the Government Cloud, a process for which Microsoft provides documentation [7], though it is a bit complicated.

Notably, other cloud providers also offer encrypted disks. In particular, Amazon Web Services' Elastic Block Storage offers encryption has a built in feature even in their government offering. Similarly, IBM's cloud platform recently introduced the feature as well.

Here enterprise management is actually relatively straightforward. Since we use the District's Azure account, the systems administrators for the Azure subscription can manage our encryption keys and audit our systems for compliance. This is one of the great advantages of cloud-based solutions for enterprises: the centralization of resource management. The Lab, in particular, has found this particular aspect of cloud computing to be extremely useful.

3.3 Secure Authentication

Finally, we come to verifying who is authorized to view data. This problem, usually called *authentication*, is quite broad as it includes how you access the secure network, firewalls, passwords, and so on. In our case, the District already has quite a few policies and quite a few pieces of software that help us maintain good practices in this area.

In this section, we enumerate a few of the best practices that The Lab employs as part of our daily operations in securing our authentication mechanisms.

3.3.1 Password Policies. First, it is important to have a good password policy for critical passwords. The Lab complies with the District's password policy [14], as maintained by OCTO, for all of our critical passwords. Key features of this policy include:

- Passwords must be at least 8 characters long,
- Passwords must contain a variety of characters, including symbols such as &, <, and >, and
- Passwords must be changed every 6 months at the latest.

3.3.2 Password Management. In addition to requiring a minimum amount of complexity in our main passwords, we also encourage the use of different, random passwords for every service we sign up for. Of course, having a different, random password for all such services would be impossible, as you could never remember all of them. As such, we use a password manager which solves this problem for us.

Having a password manager also gives us another huge advantage: the ability to *share* credentials between members of The Lab. We have several shared accounts that we manage, including our Azure accounts, and making sure that *The Lab* and not any individual Lab member owns the credentials is critical to the long-term health of the organization. Moreover, such services allow us to securely share credentials instead of simply emailing them, a highly insecure method of sharing critical information.

There are several password managers available, including 1Password[11], Dashlane[12], and Keeper[13]. Wikipedia has a long list including some open source options. After careful consideration of The Lab's particular needs, we chose to use LastPass for this purpose.

3.3.3 Two-Factor Authentication where Possible. Of course, passwords can be stolen, phished,[14] or otherwise lost. They are just strings of letter and numbers after all. A popular way to increase the security of accounts secured by these passwords is through two-factor authentication. This is a method whereby to log into a service, a simple password is insufficient. You also have to have a physical object which transmits a one-time code to you that verifies who you are. Most commonly, the physical object is your smartphone, which will typically be sent a text message with a code to type in after you have provided your password.

For all the accounts The Lab controls, including our LastPass and GitHub accounts, we require our members to use two-factor authentication. There are many introductions to two-factor authentication, e.g., [6, 15], as well as several on the pitfalls of the method [1]. While it is instructive to know the pitfalls–most prominently, avoid two-factor via text message if possible–two-factor authentication is still an important part of security best practices.

3.3.4 Wifi Connections. In general, it is good advice not to connect to open wifi routers (ones that don't say `secure` next to them or that don't have a little lock icon next to their names). Hackers can get up to all sorts of mischief if you do, especially if you start doing things like accessing your bank account. However, even connecting to a secure wifi connection can be problematic. For instance, one of the early schemes for securing wifi connections, known as WEP has been broken for over 15 years [8] with many practical attacks available by simply searching Google. Still, there are some networks that use it, and many operating systems will list such networks as "secure".

As such, we require Lab members to take the following precautions:

- When accessing Lab resources from networks that the District government does not own, always use the District's Virtual Private Network;
- Avoid accessing highly sensitive data, even over the District's VPN, while not on a home network or the District's physical network; and
- Use a strong security mechanism, such as WPA2, on home wifi networks, and avoid insecure mechanisms like WEP.

[10] When we asked Microsoft when this feature would be available in the Government Cloud, they indicated that they expected it to arrive very soon.

[11] https://1password.com
[12] https://dashlane.com
[13] https://keepersecurity.com
[14] Phishing refers to a particular kind of attack whereby someone is duped into sharing their password with nefarious actors. Perhaps the most well-known phishing attacks are carried out through email, where a seemingly legitimate email, supposedly from a trusted source like a friend or IT administrator, requests that you type your password into a website gotten to by following a link. Perhaps the most famous example of such an attack was on John Podesta, Chairman of Hillary Clinton's 2016 campaign for President of the United States [18].

3.3.5 VPNs. Virtual Private Networks are used to allow members of an organization to securely connect to protected resources when away from the office. These tools are critical for organizations with remote workers, but also make working from home or field work possible. The District maintains a VPN with various security features, including two-factor authentication for sign on and making sure that installed software such as operating systems and antivirus software is sufficiently up-to-date.

However, setting up a VPN can be extremely difficult, so The Lab has been lucky to have the help of OCTO to implement our VPN usage requirements.

3.4 A quick review

While the list above may seem long, there are really three principles that underlie The Lab's security policy: data should be encrypted in transit and encrypted at rest, and authentication to access the data should be as secure as possible.

This policy has proved satisfactory for several agencies with extremely rigorous security needs, such as the Metropolitan Police Department and the Criminal Code Reform Commission. We believe that it is a good set of practices to follow for any organization that works with sensitive data. Though, like all cyber security initiatives today, we are aware that further improvements will need to be made as best practices shift.

4 THE DATA USE AGREEMENT

When The Lab first started working in earnest with agencies at the beginning of 2017, we expected that each agency would be experienced in sharing data with outside groups, and so they would have developed standard data use agreements outlining the responsibilities those using the data must accept. As we began to work with agencies, we realized that this was often not the case. Data sharing by agencies, especially for protected data at Level 3 or above, was mostly done on an *ad hoc* basis, and also not that frequently. Since data sharing is fundamental to the operations of The Lab, it made sense for us to develop a template data use agreement which we would offer as an outline to any agency we worked with.

The document was built to allow members of The Lab and agency staff to fill in a few blanks describing the data to be shared as well as describing the project requiring the data. This sets an expectation from the legal team for what information is necessary from the technical staff without involving a long back-and-forth.

The rest of the agreement consists of language that covers the following topic areas:

- The Lab's plan to protect the agency's data, which amounts to the security policy described in Section 3;
- Procedures to follow for authorized disclosures of data as well as procedures for mitigating unauthorized disclosures of data;
- Procedures for handling unexpected, but required, disclosures of data, for instance, due to FOIA requests;
- The lifecycle of the data to be shared with The Lab, including the method and frequency of data transfers, the publication and archiving of data for replication purposes, and the deletion of data after the project ends; and
- How to modify the agreement if more data requests are necessary after the agreement has been executed.

This lays out The Lab's commitment to protecting partner agencies' data upfront while also making clear the expectation to share its results with the public.

Since we created the agreement, we have used it as the basis for agreements with the Metropolitan Police Department, the Department of Energy and Environment, the Department of Employment Services, and others. Of course, the final versions of all of these agreements often differ slightly in their details, but the broad outlines represented by the template agreement allow The Lab and partner agencies to more quickly begin negotiations and execute agreements.

5 ONGOING WORK

While these three components have greatly improved our ability to share data between The Lab and the rest of District government, there are still many hurdles we face.

5.1 The Data Policy

Perhaps most critically, the newness of the data policy is the largest hurdle. Agencies are still trying to understand how the requirements and personnel it codifies will work with existing structures. For instance, one issue the policy contemplates is that some agencies are extremely small and may not have the resources to designate an information security officer or a data officer. In this case, the policy allows the agency to outsource these positions to the office of the Chief Information Security Officer and the Chief Data Officer, respectively.

Another problem, the contours of which are not yet clear, is resolving ambiguity around the security levels attached to data. As we know from the federal government, over-classification is an impediment to the sharing of government records which would otherwise be useful to the public. As written, the data policy makes it very tempting to classify any data set as Level 3, justifying it by citing personally identifiable information contained in that data set. But aggregated and anonymized versions of these records *should* be public record, or at least available to the rest of the District's government agencies, as they would provide great value to the work of others.

The Lab hopes to play a role in resolving some of these issues. In particular, we frequently provide feedback to the Office of the Chief Technology Officer on the implementation of its data warehousing scheme. We also have found ourselves in the position of testing the limits of the data policy as we find more data sets that are collected that are useful to our efforts to build evidence-based policies and evaluate interventions as they are rolled out.

These efforts to improve the data policy are aided by our work building a community of practice of social and data science within the District's government. While some of our strategies for building this community of practice are more indirect, such as sponsoring a lunchtime talk series and pair programming with technical staff at other agencies, The Lab also has several members who are primarily employed by other agencies. These members participate in The Lab's planning process, but spend much of their time working at the agencies they are working with. In particular, one of The Lab's

members is actually employed by OCTO and detailed for a large percentage of his time to The Lab. This allows us a direct line to the Chief Data Officer so that we can facilitate a conversation between agencies that might have concerns about the data policy's provisions and the chief implementer of the policy.

5.2 The Security Policy

The threat landscape in cybersecurity is constantly changing, and so while the security policy we described above may be satisfactory at the present moment, it will eventually become out of date. The Lab plans to work with the Office of the Chief Technology Officer to make sure that the policy reflects the latest best practices.

6 CONCLUSION

Doing data science or, really, any social science requires access to data. Of course, administrative data is often quite sensitive, and so all parties to sharing the data should feel confident that the data is protected and that the planned use of the data is within the scope of how it should be used. The Lab has found that these conversations are often some of the most difficult parts of working with agencies. So we developed the documents and policies discussed above.

While the reader can find copies of the current versions of the above documents on our GitHub page[15], we also recognize that over time these documents will need to be updated. We welcome any feedback from other jurisdictions about these policies, and also conversations on our GitHub page. We hope that other jurisdictions will find these documents useful for alleviating some of the same issues that The Lab encountered as it began operations earlier this year.

7 ACKNOWLEDGMENTS

We would like to thank the many people who have developed the policies and procedures described in this brief note. Barney Kruckoff and Jenny Reed were instrumental in promulgating the new District data policy. Sam Quinney, Barry Kreisworth, and Kenneth Liebowitz took the lead on putting together the template data use agreement. And Michael South was an excellent resource for developing security best practices in The Lab.

[15] https://github.com/thelabdc/PAPER-DSSG-2017

REFERENCES

[1] Russell Brandom. 2017. Two-factor authentication is a mess. *The Verge* (July 2017). https://www.theverge.com/2017/7/10/15946642/two-factor-authentication-online-security-mess

[2] Creative Commons. 2002. Creative Commons CC0 1.0 Universal License. (2002). https://creativecommons.org/publicdomain/zero/1.0/legalcode

[3] Congress of the United States. 1967. Freedom of Information Act. (1967). 5 USC §552.

[4] Congress of the United States. 1974. Family Educational Rights and Privacy Act. (1974). 20 USC §1232g.

[5] Congress of the United States. 1996. Health Insurance Portability and Accountability Act. (1996). Public Law 104–191.

[6] Microsoft Corporation. 2017. About two-step verification. (May 2017). https://support.microsoft.com/en-ca/help/12408

[7] Microsoft Corporation. 2017. How to encrypt virtual disks on a Linux VM. (July 2017). https://docs.microsoft.com/en-us/azure/virtual-machines/linux/encrypt-disks

[8] Scott Fluhrer, Itsik Mantin, and Adi Shamir. 2001. Weaknesses in the Key Scheduling Algorithm of RC4. *International Workshop on Selected Areas in Cryptography* 2259 (2001), 1–24.

[9] Andy Greenberg. 2014. Why the security of USB is fundamentally broken. *Wired* (July 2014). https://www.wired.com/2014/07/usb-security/

[10] Mayor of the District of Columbia. 2017. District of Columbia Data Policy. (2017). https://octo.dc.gov/page/district-columbia-data-policy Mayor's Order 2017-115.

[11] Teresa McCallion. 2014. Report Card From Reno. *Integrated Healthcare Executive* (Nov. 2014). http://www.ihexecutive.com/patient-care/clinical-pathways/article/12003224/remsa-innovation-grant-results

[12] Mikaela Lefrak. 2016. To Help Combat Truancy, D.C. Schools Are Turning To A Smartphone App. (Oct. 2016). http://wamu.org/story/16/10/21/how_dc_schools_are_using_an_app_to_combat_student_truancy/

[13] Office of the Chief Financial Officer of the District of Columbia. 2017. Integrated Tax System Public Extract. (July 2017). http://opendata.dc.gov/datasets/integrated-tax-system-public-extract

[14] Office of the Chief Technology Officer of the District of Columbia. 2013. Password Management Policy. (Dec. 2013). https://octo.dc.gov/sites/default/files/dc/sites/octo/publication/attachments/2013-PasswordManagementPolicy.pdf Policy Number OCTO-PM-P101.01.

[15] SETH ROSENBLATT and JASON CIPRIANi. 2015. Two-factor authentication: What you need to know (FAQ). *CNET* (June 2015). https://www.cnet.com/news/two-factor-authentication-what-you-need-to-know-faq/

[16] Stand Out Show Up. 2015. Who We Are. (2015). http://www.showupstandout.org/who-we-are/

[17] SunlightFoundation. 2017. Open Data Policy Collection. (2017). http://www.opendatapolicies.org

[18] Joe Uchill. 2016. Typo led to Podesta email hack: report. *The Hill* (Dec. 2016). http://thehill.com/policy/cybersecurity/310234-typo-may-have-caused-podesta-email-hack