FINAL REPORT (ADC-DPT-23-N001)

January 2024

Data Protection Toolkit (DPT): A Use Case Analysis

Presented by:
NORC at the University of
Chicago

Presented to:
National Center for Science
and Engineering Statistics

Table of Contents

The America's DataHub Consortium (ADC), a public-private partnership is being utilized to implement research opportunities that support the strategic objectives of the National Center for Science and Engineering Statistics (NCSES) within the U.S. National Science Foundation (NSF). This report documents research funded through the ADC and is being shared to inform interested parties of ongoing activities and to encourage further discussion. Any opinions, findings, conclusions, or recommendations expressed in this report do not necessarily reflect the views of NCSES or NSF. Please send questions to ncsesweb@nsf.gov. The OMB control number for this collection is 3145-0215. This product has been reviewed for unauthorized disclosure of confidential information under NCSES-DRN24-004.

Introduction and Methods

Purpose

As part of the National Secure Data Service Demonstration (NSDS-D) project, the National Center for Science and Engineering Statistics (NCSES) contracted with NORC at the University of Chicago (NORC) to implement a use case analysis of the Federal Committee on Statistical Methodology's Data Protection Toolkit (DPT). The purpose of this study is to identify successful uses and potential enhancements to the Toolkit, which is intended to support agencies in expanding access to federal datasets while protecting data confidentiality. The report below describes current data protection practices reported by the federal and non-federal staff NORC interviewed, as well as their feedback and suggested improvements to the DPT.

Interview overview

NORC conducted 60-minute interviews with 15 individuals, nine of whom worked for federal agencies, and six of whom worked for state governments, universities, or research organizations. The interviews were conducted between September 22 and November 9, 2023. NORC and NCSES collaborated on the list of potential respondents to ensure representation across government, academia, and the private sector. The goal was to identify a wide range of data users who could share how they may have implemented the Toolkit, provide suggestions for improvements, and offer ideas for additional content.

Respondents represented the following types of agencies and organizations:

Agency Type	Additional Information
Principal Statistical Agency	Large and small in size
Federal Agencies without Principal Statistical Agencies	Those that fall under the Chief Financial Officers (CFO) Act and those that do not fall under the CFO Act
State/Local Data Producers and Users	---
Academic Institutions	---
Data Producer and Research Organizations	---

NCSES initiated outreach by email (Appendix A) and NORC followed up, also by email (Appendix B), to schedule the virtual interviews. NORC sent one additional email to those who did not respond to the first two outreach attempts. In October, about halfway through data collection, the team revised the initial contact email (Appendix C) and identified additional stakeholders with whom to conduct outreach and follow up, according to the same protocol used with the first round of stakeholders (initial email and no more than two follow up emails).

NORC conducted semi-structured interviews using a protocol that was developed in collaboration with NCSES. The interviews were conducted on Zoom with a primary interviewer leading the questioning and the secondary interviewer taking notes and asking follow-up questions as needed. The protocol (Appendix D) collected information on the respondent's role, their organization's approach to data protection, current resources used to support data and confidentiality protection, and their experiences with the DPT. Follow-up questions were tailored depending on whether a respondent had familiarity with the DPT. If the respondent was not familiar with the DPT or had not seen the current version, NORC shared the link to the DPT and asked the respondent to look at a few preselected sections, tailored to their role. NORC then asked for any first impressions and feedback, with the understanding that it is difficult to provide immediate feedback. The remainder of this report presents an Executive Summary followed by a detailed discussion of findings, organized by topic.

Executive Summary

Purpose

As part of the National Secure Data Service Demonstration (NSDS-D) project, the National Center for Science and Engineering Statistics (NCSES) contracted with NORC at the University of Chicago (NORC) to implement a use case analysis of the Federal Committee on Statistical Methodology's Data Protection Toolkit (DPT). The purpose of this study is to identify successful uses and potential enhancements to the Toolkit, which is intended to support agencies in expanding access to federal datasets while protecting data confidentiality.

Methods

NORC conducted 60-minute interviews with 15 individuals, nine of whom worked for federal agencies, and six of whom worked for state governments, universities, and research organizations. The interviews were conducted between September 22 and November 9, 2023. NORC conducted semi-structured interviews using a protocol that was developed in collaboration with NCSES. The interviews were conducted on Zoom with a primary interviewer leading the questioning and the secondary interviewer taking notes and asking follow-up questions as needed.

Findings

The current data protection environment

Respondents from both federal agencies and non-federal organizations are aware of and invested in the need for data protection. All reported that various protocols and formal procedures are in place for making data available while preserving privacy and confidentiality. However, it is unclear how consistent these methods are across organizations, and between the federal and non-federal landscapes. Federal agencies are struggling with operationalizing Evidence Act guidelines around data protection, and all respondents are facing decisions about the use of differential privacy practices.

Key findings from the interviews:

- Federal agencies employ multiple techniques to avoid disclosure of private and confidential information when making data available to the public or researchers. A few examples include cell suppression, variable suppression, or noise infusion.
- Federal agencies interviewed used a tiered access model for providing data access to external researchers.
- Federal agencies maintain strict protocols for sharing linked data.
- Two respondents noted that their data were never linked due to the siloed nature of their agency's program offices.
- Non-federal organizations interviewed also employ techniques to avoid disclosure of private and confidential data.
- Current variable suppression protocols allow reviewed data to be publicly released, though sensitive data is still available by request.
- All federal agencies interviewed have a disclosure review process for all publicly released data. Agencies use a variety of resources to guide their review process. Some agencies interviewed have a Disclosure Review Board (DRB) to support the process.
- Non-federal staff follow formal disclosure risk review guidelines and described the resources used to guide their disclosure review process.
- Non-federal respondents described their processes to control access to sensitive data and ensure that output is void of confidential information.
- Federal agency respondents described confusion surrounding current data protection initiatives tied to the Evidence Act.
- Respondents from state governments noted the importance of meeting both state legislative requirements around data release and data confidentiality and federal collection requirements for state-based federal surveys.
- Federal respondents said they work to ensure confidentiality practices are consistent from the point of data collection to data release.
- The federal agencies interviewed are balancing the desire to make more data available to the public, while protecting confidential information.

- The usability of demographic variables depends on whether the data is being used for research or for program monitoring and administration.
- Some federal agencies are working to implement differential privacy practices.
- Non-federal respondents expressed concern about differential privacy practices, perceiving them as unnecessary and potentially reducing data usability. Many non-federal respondents were focused instead on decreasing response rates in government surveys and the need to explain to small communities and minority groups why it is important they participate.

Feedback on the Data Protection Toolkit (DPT)

None of the respondents who participated in the Use Case Analysis interviews reported using the DPT as a regular resource. When asked directly about incorporating it into their daily data protection practices, some responded they would probably continue to turn to their trusted colleagues rather than start using the DPT, while others saw that it could be a helpful reference, particularly to provide clear language defining data protection terms and concepts. When imagining how the DPT might be made more useful, respondents suggested adding information about new data concepts such as differential privacy, improving the navigability of the Toolkit, and including guidance directed more explicitly to non-federal organizations.

Key findings from the interviews:

- Federal agency staff thought that the Toolkit looked like a helpful resource potential source for standardized information were all federal agencies to use it consistently.
- Respondents envisioned using the Toolkit to learn and understand new data concepts such as differential privacy.
- One respondent was using resources in the Toolkit to support their work in establishing a DRB at their agency.
- Some federal agency respondents were less sure if they would use the Toolkit as a reference because they knew the field and would turn to their colleagues with any questions.
- Respondents wondered if there would be additional information in the Toolkit on different data types.
- Non-federal staff noted that some of their organizations had disclosure review boards, but that the information on DRBs in the Toolkit was specific to federal DRBs.
- Respondents thought that the resource list in the Toolkit was extensive and that it would take time to comb through it to access those most relevant to a particular situation or skillset.
- Federal agency respondents provided additional suggestions on improving the usefulness of the Toolkit, including adding a section on assessing data risk and providing resources about unsuccessful data protection efforts and lessons learned.
- Respondents noted that although there were different sections of the Toolkit, the intended audience and knowledge level were not explicit.
- Non-federal respondents suggested methods to introduce the Toolkit to non-federal audiences, such as through live or recorded webinars and completion certificates.

Detailed Findings

None of the interview respondents were active users of the Toolkit. Some had heard of it through professional meetings or had been asked to contribute resources or writings to it, but none were actively using it as a resource. The interviews produced a wealth of information about current resources and practices that are in place at each organization. The findings below focus first on those topics, as well as on current concerns about data protection, and data usability. This section concludes with feedback specific to the Toolkit, including what respondents perceived as helpful elements of the Toolkit along with their suggestions and recommendations.

The current data protection environment

Data protection practices

Federal agencies employ multiple techniques to avoid disclosure of private and confidential information when disclosing or releasing data. These techniques include using cell suppression (suppressing values in a table with small counts), secondary suppressions (suppressing additional cells that do not have small counts but could be used to determine the values of suppressed data), and variable suppression (suppressing an entire variable). Variables that are suppressed include personal identifying information (PII), geographic information (addresses, county, etc.), and any other variables that could facilitate identification of individuals. In addition, data are often not released at an individual or household level to protect confidentiality. When household-level information is released, agencies are careful to meet minimum requirements for reporting household counts and ensure numbers are sufficiently large enough that the risk of a mosaic effect (using combined datasets to figure out new information) when linking data is determined to be low.

In addition to suppressing variables and cells that could lead to reidentification of respondents, a few respondents explained that their agencies also used additional measures for privacy protection, specifically mentioning data perturbation and data coarsening techniques (adding noise to their data). Noise infusion includes rounding values and adding random values of a selected variable to protect respondents' data from being disclosed.

Federal agency data sharing

Federal agencies interviewed used a tiered access model for providing data access to external researchers. While agencies previously exclusively used data enclaves to provide access to data for approved researchers, now outside researchers go through the Standard Application Process (SAP) to apply for access to confidential data from the 16 principal statistical agencies and units. Agencies interviewed confirmed their understanding that now the SAP is the only way to apply for access to restricted data from federal statistical agencies for research and statistical purposes. Once the application is approved, respondents reported that individual agency staff manage the related procedures and processes for accessing the data, which may include signed Memorandum of Understanding (MOU) detailing the responsibilities of the institution and the researcher in preventing data breaches or inadvertent disclosure. One respondent in the federal government had questions about tiered access because they struggled with the “notion that a researcher is coming to us and wanting to use data and the agencies have to provide a lot of help.” They had head discussions about agencies being “more proactive and offering services…rather than someone being smart enough to run their own programs.”

Federal agencies maintain strict protocols for sharing linked data. Linked datasets are highly restricted and generally only released in the form of aggregate statistics. One agency explained that when they collaborate with another agency or external organization, they set up restrictions so that each agency or organization had only the information they needed, and no entity had access to all of the information that could be triangulated to reveal private information.

Two respondents noted that their data were never linked due to the siloed nature of their agency’s program offices. These respondents said their agencies are working on building systems to integrate the data collected from their program offices so that researchers at the agency could access the data across the agency rather than submitting data requests to each program officer.

Non-federal data protection practices

Non-federal organizations interviewed also employ techniques to avoid disclosure of private and confidential data. Due to smaller sample sizes, non-federal organizations reported suppressing most of their granular geographic data. Data below the county level are frequently not publicly available due to concern that households could be reidentified by combining multiple data sources containing demographic and lower-level geographic data. For example, they suppressed zip code and county data, as these are considered potentially identifiable when combined with other demographic data, such as occupation (particularly for rural counties).

All non-federal participants described similar organizational protocols to ensure any publicly released data could not be used to identify households or individuals. Non-federal agencies and organizations said that they have created internal resources describing specific protocols and data preparation techniques. They work collaboratively to share information about best practices for facilitating confidentiality and mentor new staff members to ensure that the data protection protocols are understood and implemented. Respondents from state governments noted looking at the practices of other states and at the research other states had pulled together as guidance for best practices. They pointed out that most research and coding guidance was on state-level data and they “hadn’t found anything that was specific to smaller levels of geography like the county level.”

Current variable suppression protocols allow data to be publicly released, though sensitive data is still available by request. Non-federal respondents noted that almost all of their data at the state-level was published in data tables on public data platforms and websites. The aggregate data tables that are publicly available contain the information that is most requested and believed to be most useful. Upon request, states also share additional data variables and data tabulations that are not in the public use database, as long as the data do not contain identifying information and researchers sign a Data Use Agreement (DUA) or MOU. Data is then provided via encrypted download or a data enclave. One non-federal respondent noted that it is important to educate researchers who request data beyond what is publicly available because often that restricted data is not needed to answer the research question.

Disclosure review resources

Federal disclosure review resources

All federal agencies interviewed have a disclosure review process for all publicly released data. This process depends on the size of the data set used in the research. Two agencies described additional tests for large data sets, including dominance and sensitivity tests for published tables (testing for the superiority of a certain statistical model and testing the uncertainty in output given different input variables), and having different rules for weighted and unweighted data. Federal staff interviewed explained that there are processes in place to review estimates for privacy and reidentification concerns before release, and to ensure that released estimates meet the requirements of the Confidential Information Protection and Statistical Efficiency Act (CIPSEA). Federal staff also review research presentations and papers to make sure estimates cannot be used to reidentify respondents. One agency described how disclosure limitation methods were codified in a large manual so that everyone is following the same procedures around maintaining confidentiality. When new employees join, they are given confidentiality training on the internal procedures and policies surrounding access to confidential data as well as a written standard operating procedure for each dataset. Another agency noted that they revise their policies every couple of years "depending on evaluation that this program is not working very well" to produce the best data products. When respondents discussed sharing data with another agency, they noted that the disclosure review process had to take place in both agencies. Not only do both agencies want a say in how the data are released, but both are also responsible for signing off on the publicly released data.

Agencies use a variety of resources to guide their review process. Federal staff described the Five Safes (Safe People, Safe Projects, Safe Settings, Safe Data, and Safe Output) as a useful paradigm for assessing the risks of disclosure. Respondents also relied on resources such as Statistical Policy Working Paper 22 (Report on Statistical Disclosure Limitation Methodology[1]), conference proceedings, or journal articles. They reported that some agencies share code with one another to assist with preparing data for disclosure and ensuring that it meets disclosure policies.

[1] https://www.fcsm.gov/assets/files/docs/spwp22WithFrontNote.pdf

Some agencies interviewed have a Disclosure Review Board (DRB) to support the disclosure review process. The DRBs are comprised of statisticians, economists, legal experts, and policy officials and are responsible for ensuring that agency procedures are followed. They will review any special techniques or procedures used to ensure that the released estimates protect confidentiality. On occasion, if there are concerns with the analysis plan submitted via the SAP, DRB's are consulted on data questions and confidentiality concerns before a researcher begins their work. In all cases, DRBs are engaged at the end of the research project to monitor output and ensure that it does not contain confidential information. The DRBs also work with program offices to improve their disclosure avoidance procedures and make changes as techniques evolve. Federal agency respondents discussed how agency size determines whether certain data protection resources are available to support disclosure avoidance. One respondent commented that they could imagine an "organization even smaller than us having just been completely paralyzed by the idea of releasing something that they shouldn't be releasing because they don't have the resources." These agencies may not have the resources to staff a DRB with data privacy experts but may instead rely on staff with other skill sets to review data being released.

Non-federal disclosure review resources

Non-federal staff follow formal disclosure risk review guidelines and described the resources used to guide their disclosure review process. Non-federal staff said they learn disclosure review procedures from several sources including the knowledge and experience of their colleagues. Several reported using federal survey documentation as a reference for learning about protecting data confidentiality. Respondents also reported learning by talking to peers at conferences, reading journal articles, and by reading the Institutional Review Board (IRB) websites of universities and organizations. Non-federal respondents expressed a need for education and guidance on how to release data while being considerate of confidentiality concerns. One respondent explained that some individuals have responded to data protection protocols by believing that "we can't release anything. That's not the case. We can release things, we just have to be careful, whereas some people are like 'why can't we release everything?'"

Non-federal respondents described their processes to control access to sensitive data and ensure that output is void of confidential information. Non-federal staff at universities reported having DRBs staffed with data security officers, data privacy officers, members from the security team, and outside disclosure risk experts. In places where there is not a formal DRB, the staff who oversee the creation of public use and restricted use data files also oversee data access and disclosure. Virtual data enclaves and physical data enclaves are also used for data access. One respondent described an encrypted download option that researchers can receive after filling out a data security plan, to access data that is not publicly available. The output is then vetted before being given back to the researcher to ensure that it does not contain confidential information. They review the data for direct and indirect identifiers, particularly those in open-text responses.

Data protection protocols

Federal agency respondents described their response to the Evidence Act and confusion surrounding current data protection initiatives. Respondents from non-statistical agencies discussed efforts to create offices devoted to data confidentiality as a response to the Evidence Act. The number of both agency-led and federal government-led initiatives around data and the Evidence Act make the efforts challenging to understand and distinguish. There was some concern that agencies had different interpretations of the Act. Another respondent noted that there was work going on in the Office of Management and Budget (OMB) and the Interagency Council on Statistical Policy (ICSP) around the regulations for tiered access[2] and an outline has been provided to statistical agencies for review. One respondent was concerned that regulations and resources tended to only consider the risk of the individual and did not include language about the “risk to the reputation of the agency” with a data breech and the “risk of the agency to be able to collect” confidential information.

There was some confusion about the different roles of America’s DataHub Consortium (ADC), Federal Statistical Research Data Centers (FSRDCs), the work of the individual agencies, and the Advisory Committee on Data for Evidence Building (ACDEB), as it seemed that each “are going off in their own directions.”

Respondents from state governments noted the importance of meeting both state legislative requirements around data release and data confidentiality and federal collection requirements for the state-based federal surveys. Federal agencies have been encouraging and funding the states to publish more of their data. One agency discussed how their data are collected through the states and much of it is released by the states. When asked about the role of a potential national secure data service, respondents expressed concern about how it would engage federal, state, local, and tribal governments in working together and having consistent practices by using the “same disclosure procedures.”

Data protection concerns and data usability

Federal data quality

Ensuring confidentiality practices are consistent from the point of data collection to data release builds trust both with the individuals and entities responding to federal data collections and those using the data. The federal respondents who made this point also noted that program offices are often siloed from each other, resulting in varying data management practices. One federal respondent also noted the importance of coordinating confidentiality processes across program offices to ensure high quality and consistent compliance.

[2] https://www.whitehouse.gov/wp-content/uploads/2022/12/M-23-04.pdf

The federal agencies interviewed are balancing the desire to make more data available to the public, while protecting confidential information. Respondents expressed concerns about balancing data protection with usefulness. More specifically, federal staff noted the need for the data to be accurate, timely, and relevant to their audience. There was consensus that agencies want to provide both informative data, while maintaining confidentiality protection. It is important that the federal data produced is high quality so that researchers and the general public are not making conclusions based on misleading data. One respondent noted that they would like to do more data linkage work and produce more aggregated data but were not sure how to go about doing this. Another agency mentioned that they had recently reviewed their data products to determine compliance with the Evidence Act. After completing the review, they removed some publicly available data because it did not meet the agency's quality standards. These data products will be revised and then reposted.

The usability of demographic variables depends on whether the data is being used for research or for program monitoring and administration. Two respondents mentioned the focus on racial equity of the current administration and the new lens that puts on the usability of data products. For example, though data on race and ethnicity are collected, federal programs do not use race and ethnicity to determine qualification. Therefore, race and ethnicity variables are not critical as they did not play a role in program eligibility and monitoring

Differential privacy concerns

Some federal agencies are working to implement differential privacy practices. One agency noted that they were working with an external group to study the feasibility of differential privacy and that the external group had agreements with other agencies as well. One respondent expressed concern about differential privacy extending the time it takes to publish a data product.

These new differential privacy practices were concerning to non-federal respondents who tended to perceive them as unnecessary and potentially reducing data usability. Some respondents were concerned that differential privacy and synthetic data would prevent the measuring of small populations and the release of data that state and local governments need to allocate resources, administer services and programs, and understand their populations. Non-federal respondents described their need for a trustworthy benchmark to use to allocate resources, particularly for vulnerable groups like undocumented immigrants who may not appear in other official statistics. In particular, the U.S. Census Bureau is the go-to place for reliable microdata and researchers and state officials are concerned about a new focus on aggregate data delivery. In addition, due to noise added, different data releases from the same census may have different numbers, thereby increasing the difficulty in relying on the numbers and using the data to understand their communities. One respondent explained, "there are folks now who are saying, well, if I can't get the data that I need at this level because of confidentiality should I just be real relying on Google or somebody else to get my information? Sure, we don't know how good that data is, but it's better than no data. And then with the noise that's being added to it, how much confidence do we have in census data?"

Many non-federal respondents noted decreasing response rates in government surveys and the need to explain to small communities and minority groups why it is important they participate. While individuals wanted their identity and confidentiality respected, they also wanted to be able to see themselves in the data. Several non-federal respondents expressed that the decennial census is being advertised as a way to provide data and information to allocate resources to communities, but the data is not released at a level small enough to do that. There were frustrations with small communities not having information about themselves to advocate for their own communities.

Feedback on the Data Protection Toolkit

Aspects of the Toolkit that appeared helpful

Though no respondents were active users of the DPT, all federal respondents were familiar with it from professional working groups, meetings, and their colleagues. One federal staff member had done presentations on the Evidence Act, which included telling people about the DPT to encourage agencies to look beyond their own internal resources, though he had not used the DPT in his day-to-day work. All respondents were given the link to the Toolkit during the interview, provided with an opportunity to review a few sections relevant to their work, then asked for feedback on those sections. A few respondents had looked at the Toolkit in preparation for the interview. All were eager to hear that it was publicly available and that they could share the resource with their colleagues.

Federal agency staff thought that the Toolkit looked like a helpful resource, noting that if federal agencies were to use it more actively, it might help standardize data protection practices across agencies.
The Toolkit provides a central location for resources, making them easier to access and share with others. While looking at Toolkit resources about data enclaves and tiered access, one respondent liked how all resources were in one place and how easy it is to share the links to the resources with colleagues. Another federal respondent who is working with state agencies on strengthening their data analysis practices resolved to share the Toolkit with states as part of that effort. Respondents saw it as a useful tool to assist with teaching others about data protection and educating staff in their agencies. They noted it may be particularly useful to smaller agencies and administrations that may not have their own internal resources, particularly the sections on assessing disclosure risk and understanding the SAP. Respondents also imagined the DPT could be used as a reference for published papers and to cite methodologies.

Respondents envisioned using the Toolkit to learn and understand new data concepts such as differential privacy. Most respondents saw that the Toolkit could create a shared understanding of these concepts across the federal agencies. The accessible language of the Toolkit made it easier to communicate the importance of data protection with leadership. The Toolkit helped another respondent connect concepts such as disclosure risk, disclosure avoidance processes, and tiered access, and understand that some data that is broader can be made more easily available than data that is more likely to reveal confidential information. Respondents also liked that the Toolkit provides resources on sharing data while protecting confidentiality, such as the virtual data enclave section on sharing federal agency data with researchers. One respondent who explored the training modules described them as interactive and engaging and thought it would be helpful for the trainings to end with suggestions of resources in the Toolkit to look at next.

One respondent was using resources in the Toolkit to support their work in establishing a DRB at their agency. This respondent appreciated the available documents, but was having some difficulty understanding which DRB example charters were most relevant to their agency. They suggested that it would be helpful to have an explanation of how the examples differ, such as by type of data or size of agency, so that people can understand which of the charters would be the most useful to reference.

Opportunities to improve the Toolkit

Some federal agency respondents were less sure if they would use the Toolkit as a reference because they knew the field and would turn to their colleagues with any questions. One respondent wondered if there would be newer updates or information on newer topics such as the impact of artificial intelligence on the mosaic effect.

Respondents wondered if there would be additional information in the Toolkit on different data types. Several respondents noted that most disclosure avoidance activities focus on individuals and households. However, they thought it would also be helpful to have information pertaining to protecting confidentiality when the unit of the data was different, such as employers within establishment data. They appreciated seeing examples of best practices from other federal agencies in the Toolkit and would like to see examples of best practices for different data types.

Non-federal staff noted that some of their organizations had disclosure review boards but that the information on DRBs in the Toolkit was specific to federal DRBs. These respondents felt that additional information and examples of what effective DRBs look like outside the federal context is missing from the Toolkit.

Respondent recommendations for improving usability of the Toolkit

Respondents thought that the resource list in the Toolkit was extensive and that it would take time to comb through it to access those most relevant to a particular situation or skillset. It would be helpful to include a navigation page or display topics and subtopics on a side panel so that users can find information as the Toolkit grows. A site map or other navigation page would also help with understanding the source of any information found on the website. Other comments included the need to have a good glossary so that there were standard definitions for data protection terms.

Federal agency respondents provided additional suggestions on improving the usefulness of the Toolkit. One agency had an idea for a validation server where researchers could submit code and the agency would run the code for the researcher. The output would tell them how close the actual answers were from the publicly released synthetic data—though this would require additional resources and support on behalf of the agency. Federal staff are interested in seeing a section of the Toolkit explain how other federal agencies assess data risk. They also said it would be helpful to have resources on things that had been tried but did not work so that others did not repeat similar endeavors.

Respondents noted that although there were different sections of the Toolkit, the intended audience and knowledge level were not explicit. One respondent suggested differentiating resources by intended audience or level (such as with a symbol). While data protection experts know the information they are seeking, for someone who is less familiar with the topic it is difficult to determine which information is most pertinent. This is particularly relevant in the readings section where a search may pull up many readings, but it is hard to differentiate the level and targeted audience for each. The resources are currently organized by topics, but it would be helpful if they could be sorted with other filters, such as knowledge level or type of organization. One respondent also noted that the search function pulled up sections called “Recommended Resources” and “Other Resources.” He wondered about the different between the two, and how a resource would come to be designated as “recommended” versus “other.”

Non-federal respondents suggested methods to introduce the Toolkit to non-federal audiences. They thought that introducing the Toolkit through webinars would be helpful because the webinars could be tailored to different audiences. It would force people to carve out the time to learn about the Toolkit from an expert. Two respondents suggested offering some sort of completion certificate with the tutorials online so that people could show that they had taken the training on Toolkit guidelines. The certification would serve as an additional endorsement for the best practices.

Discussion

Summary of recommendations

Toolkit organization

After their review of the DPT, respondents provided recommendations for the Toolkit. Regarding the organization of the online Toolkit, respondents suggested improving existing navigation and definitional tools, such as adding a site map and glossary for standard tools and adding refined searching capabilities in the resource section. Respondents also suggested adding the ability to filter the search for resources by user experience (new to data protection versus being an expert in it), and creating icons that help differentiate information in other areas of the website by intended audience and/or user experience level. Finally, there were requests to add additional contextual information to the DRB examples to understand which ones were for large versus small agencies, and which would work best in a non-federal context.

Communication and education

Respondents were happy to learn that the Toolkit was publicly available, but while many had heard about it, they felt they had not been informed when it went "live." To better publicize the existence of this resource, respondents suggested live webinars to publicize the Toolkit and explain its features and uses. This way users would have dedicated time to learn from and ask questions of the Toolkit designers. Recordings of the live webinars could then be added to the DPT website as an available reference. Others suggested the creation of a distribution list to provide alerts when the Toolkit is updated with new information and additional resources. Finally, some respondents offered the idea of formalizing any training around the DPT with completion certificates. Non-federal staff felt such certificates would particularly be helpful in showing they had taken the time to learn the federal standards around data protection and disclosure.

Suggestions for additional information on the Toolkit

The interviews produced several suggestions for information that could be added to the Toolkit. Respondents were curious about "failed" data protection endeavors and were interested in hearing examples of processes and methods that agencies had tried but that ended up not working. Respondents were also curious about the best resources to look at after completing the interactive training modules and wondered if there could be a section at the end of the module of suggested resources to look at next. There also was a suggestion to have resources on protecting establishment data, as most resources seemed to be on protecting the privacy of individuals. Finally, non-federal respondents requested an example of a non-federal DRB charter and guidance on how to implement a DRB in a non-federal context.

Conclusion

Interview conversations and findings reveal a vivid interest from both federal and non-federal staff in learning about and properly implementing strong data protections that still provide access to accurate data for evidence-building research. The DPT is perceived as a valuable resource but is not utilized to its fullest potential by its intended audience. In addition to the respondent suggestions described in previous sections of this report, we offer two additional ideas for expanding awareness and use of the Toolkit. First, given that many respondents reported reaching out to knowledgeable people within and outside their organizations for help with specific questions and issues, consider facilitating those kinds of conversations so that more staff can tap into them, or capture and provide their content in an organized way within the DPT (e.g., enhanced or super FAQs). In addition, non-federal respondents are modeling their data protection practices after federal guidelines, but feel the DPT does not explicitly acknowledge where there may be legitimate differences for non-federal entities. Perhaps there is opportunity for the DPT to fill that perceived gap and provide explicit guidance for non-federal audiences to support and strengthen their data confidentiality and disclosure practices. A combination of specific changes to the Toolkit and coordinated efforts to more widely publicize its existence and value will serve to increase its use across the community of evidence-building data users.

Appendix A: NCSES Outreach Email

Subject: Request to Participate in an Interview about Data Protection

Dear [NAME],

The National Center for Science and Engineering Statistics (NCSES) has contracted with NORC at the University of Chicago (NORC) to implement a use case analysis of the Federal Committee on Statistical Methodology's Data Protection Toolkit (DPT). The purpose of this study is to identify successful uses and potential enhancements for the toolkit that strengthen its ability to enable access to federal datasets while protecting data confidentiality. As part of this effort, NORC will be conducting key informant interviews with representatives from a variety of sectors, including government agencies, academia, and the private agency. We will use these interviews to gauge needs, inform improvements to the Toolkit, and identify any additional content for the toolkit.

You will receive an email from NORC in the next few days providing additional information about this project and requesting to set up a time to talk. We strongly encourage you to contribute your perspectives and insights as we work to enhance the toolkit and improve access to information about data disclosure and ensuring confidentiality. The more viewpoints and input we have, the more we can ensure the Toolkit meets the needs of a wide variety of users. If you have any questions about this project, please contact Heather Madray, hmadray@nsf.gov. We thank you for your consideration of this request and for your time and expertise.

SIGNATURE

Appendix B: NORC Follow-Up Email

Subject: Follow-up on NCSES Request to Participate in an Interview about Data Protection

Dear [NAME]

NCSES reached out recently about an effort to implement a use case analysis of the Data Protection Toolkit. The National Center for Science and Engineering Statistics (NCSES) has contracted with NORC at the University of Chicago (NORC) to implement a use case analysis of the Federal Committee on Statistical Methodology's Data Protection Toolkit (DPT), identifying successful uses and potential enhancements for the toolkit that strengthen its ability to enable access to federal datasets while protecting data confidentiality.

You were recommended to us by [name/organization] as someone who would provide an important perspective on this work, and we are reaching out to see if you would be interested in participating in a confidential virtual discussion with an interviewer from NORC, about your work in data protection and the resources you use. We are interested in your perspective even if you are not familiar with the DPT or have never used it. All interview responses will be kept confidential and reported in aggregate. NORC will not share your interview responses with anyone outside our study team.

We are now following up to coordinate a date and time for this interview that is convenient for you within the next week. We expect the discussion to last 90 minutes. In the table below please indicate which dates and times work for you. We will then send you a link to a meeting for one of your preferred dates. When you respond with your date and time preferences, please also note whether you prefer the interview to be conducted with or without video, as well as which videoconferencing platform you prefer.

Date	Time	Videoconferencing platform preference	With video ok? Y/N	Notes, questions, comments

We understand that you are busy and in different time zones, and we are flexible. If none of the above times work for you, please send us a time that would work for you for a 90-minute interview.

We are interested in capturing perspectives from many users and potential users. Are there other people you work with who you would recommend we reach out to?

We are happy to answer any questions or address concerns you may have about the interview's content or purpose. Please reach out to Julie Kubelka with any questions (Kubelka-julie@norc.org) We look forward to hearing from you!

Thank you,
The NORC Study Team

Appendix C: Revised Outreach Email

Dear XXX,

The National Center for Science and Engineering Statistics (NCSES) has contracted with NORC at the University of Chicago (NORC) to implement a use case analysis of the Federal Committee on Statistical Methodology's Data Protection Toolkit (DPT). The purpose of this study is to assess the potential for expanding awareness and use of the toolkit to support access to data while protecting data confidentiality. As part of this effort, NORC will be conducting key informant interviews with representatives from a variety of sectors, including federal and state government agencies, academia, and the private sector. We are interested in having a better understanding of the resources people use when they are working on issues of data protection as well as suggestions to strengthen the usability of the Data Protection Toolkit. We will use these interviews to gauge awareness, explore needs, inform improvements, and identify any additional content for the toolkit.

You will receive an email from NORC in the next few days providing additional information about this project and requesting to set up a time to talk. We strongly encourage you to contribute your perspectives and insights as we work to enhance the toolkit and improve access to information about data disclosure and ensuring confidentiality. The more viewpoints and input we have, including from those who have not used or are not yet aware of the toolkit, the more we can ensure the toolkit meets the needs of a wide variety of users. If you have any questions about this project, please contact Heather Madray, hmadray@nsf.gov. We thank you for your consideration of this request and for your time and expertise.

Heather

Appendix D: Interview Protocol

Hello. My name is [NAME], and I am a [TITLE] in the [DEPARTMENT NAME] department here at NORC.

Thank you for your willingness to participate in this interview today. We know you're busy and we really appreciate your time. Through this project we're interested in the hearing about the kinds of experiences you have had when seeking information about data protection, and any experiences you may have had using resources to support your work in assessing disclosure risk, and maintaining confidentiality when sharing data . We plan to interview about 15-20 individuals, and our goal is to use these discussions to get a better understanding of the resources people use when they are working on issues of data protection as well as suggestions to strengthen the usability of the Data Protection Toolkit.

We anticipate that this discussion will last around 60 minutes. Your participation is voluntary, and you can conclude the discussion at any time. You may also skip any questions you don't feel comfortable answering. All interview responses will be kept confidential and stored separately from your contact information, and we will not share your contact information or interview responses with anyone outside the NORC study team. When we submit our final report of findings, your responses in that report will be combined with the responses of around 20 other individuals.

Do you have any questions about your participation?

Do you agree to participate in this interview? (Yes/NO)

I will be taking notes, but we also would like to record our conversation to make sure that I don't miss anything important. The notes and recording will only be used by this study team to write our report. All audio recordings will be destroyed once the project is complete.

Do you agree to having this conversation recorded? (Yes/NO)

Ok, the recording has started. Do you have any questions or concerns before we get started?

Introduction

1. How would you describe your current role?
    - In your current role, what specific aspects do you focus on related to data disclosure and confidentiality?
    - What are you doing regarding disclosure?
2. In your role as a [policymaker, agency executive, practitioner, subject matter expert, FILL FROM RESPONSE TO #1] can you think about and describe a recent instance when you were working on making data available while protecting confidentiality?
    - What questions did you have?

- What information did you need?
- What resources did you draw upon? Why?
- What purposes did each resource serve?
- How did you originally find out about these resources?

3. Were the resources you used helpful?
    - Which resources were the most helpful? Why?
    - Which resources were the least helpful? Why?

Data Protection Toolkit

4. Have you heard of the federal Data Protection Toolkit? [IF NO, SKIP TO QUESTION 8]
5. Have you had a chance to look at the Data Protection Toolkit? [IF NO, SKIP TO QUESTION 8]

Data Protection Toolkit - Users

6. If so, what brought you to look at it? (Google, colleague reference, etc.)
    - Please describe in as much detail as you can the circumstances that led you to use the DPT.
    - What were your general reactions to the toolkit?
    - Which sections did you look through?
7. Did you end up using some of the resources you found there?
    - [IF YES] Describe in as much detail as you can how your organization used the DPT. [SUGGEST SCREEN SHARE WHERE APPROPRIATE, FOR RESPONDENT TO DEMONSTRATE USE OF DPT]
        i. What was that experience like?
        ii. What did you look for in the Toolkit and didn’t find?
        iii. In what ways did you end up using the resources?
        iv. What aspects of the Toolkit worked well?
            1. PROBE ON: Access, usability, accessibility, protections, anything else.
        v. What aspects of the Toolkit didn’t work well?
            1. PROBE ON: Access, usability, accessibility, protections, anything else.
        vi. Will you use the Toolkit again? Why or why not?
    - [IF NO] Please explain why you ultimately did not decide to use the DPT. [SKIP TO QUESTION 12]

Data Protection Toolkit – Non-Users

8. Let me share my screen so you can see the website. Let’s start by looking at this introduction page: https://nces.ed.gov/fcsm/dpt/learn or https://nces.ed.gov/fcsm/dpt/content/1

9. As a [ROLE] let’s start by looking in the resources in [SECTION]
    - As an *Agency lead*, let’s start by looking at resources in the Promoting Data Access While Protecting Confidentiality Section. The goal of this section is to provide a high-level summary of the issues and challenges of protecting data while promoting their use for research and evidence building. I’m going to give you a few minutes to poke around this section.
        i. What are your first impressions/general reactions to this section?
        ii. What is not here that you would like to see?

- As a *practitioner,* let's start by looking at resources in the Tiered Access Models section. The goal of this section is to understand approaches and solutions for providing controlled access to confidential data to support evidence building. I'm going to give you a few minutes to poke around this section.
    i. What are your first impressions/general reactions to this section?
    ii. What is not here that you would like to see?
- Next, let's look at the Statistical Disclosure Limitation section. The goal of this section is to understand the statistical techniques and tools that can be used to reduce the likelihood that individuals can be re-identified in data products. I'm going to give you a few minutes to poke around this section.
    i. What are your first impressions/general reactions to this section?
    ii. What is not here that you would like to see?

- As a *subject matter expert,* let's start by looking at the resources in the Reading Room. The goal of this section is to serve as a depository of resources collected across government agencies and to provide information that may assist anyone doing disclosure review with content specific to their roles. I'm going to give you a few minutes to poke around this section.
    i. What are your first impressions/general reactions to this section?
    ii. What is not here that you would like to see?

10. PROBE ON OTHER SPECIFIC SCREENS AS IDENTIFIED BY NCSES
    - Assessing Disclosure Risk (particularly the "other resources" section)
    - Search Function
    - Online Training Modules
11. What are your first impressions/general reactions to the toolkit?
    - Would this toolkit have been helpful and provided you with resources that you needed in the scenario you described to me earlier?

Closing Questions

12. Are there missing features on the DPT that we need to address?
13. Are there other resources we have not yet discussed that you draw upon when working on data protection? If yes, please describe the resources and how you use them in your work.
14. Are you aware of other DPT users we should include in this study? Who should we reach out to?
15. Thank you very much for your time today. Is there anything else we didn't discuss that you would like to share before we end this conversation?

I appreciate the important insights you have shared. If you have any questions or think of anything else, please reach out.