

May 2024

Environmental Scan and Outreach Report:

Findings from federal statistical
agency and data user interviews

Presented by:
NORC at the University of
Chicago

Presented to:
Heather Madray, National
Center for Science and
Engineering Statistics

Table of Contents

Introduction and Methods	1
Purpose	1
Introduction	1
Interview Methodology	1
Federal Statistical Agencies	1
Non-Federal Data Users	2
Executive Summary	3
Part 1 – Common themes in federal statistical agencies’ support for their data users	4
Data Discovery.....	5
Determining fitness for use.....	5
Availability of metadata	5
On-line inventory/portal	5
Data Access.....	6
Data accessibility	6
Gaining access.....	6
Securing legal/policy authority to access.....	7
Data Usage.....	7
Authority to publish.....	7
Ability to link and link keys available.....	7
Determining quality of linkage	8
Part 2 – Common themes in federal data users’ experiences with statistical data	8
As-is Data Discovery.....	9
Determining fitness for use.....	9
Determining availability of metadata.....	10
Reviewing online inventories	10
As-is Data Access.....	10
Experiences of data accessibility.....	10
Gaining access.....	11
Securing legal/policy authority to access	11
As-is Data Usage	11
Authority to release tabular results	11
Authority to link and link keys available	12

Determining quality of linkage	12
Part 3 – Challenges faced by data users	12
Data discovery challenges	12
Searching for data.....	12
Determining data fitness.....	13
Data access challenges	13
Navigating restricted access.....	13
Meeting training and credentialing requirements	13
Data use challenges	13
Validating statistical analyses.....	13
Linking data.....	14
Reviewing disclosure risks	14
Part 4 – Opportunities to support data users	14
Data Discovery.....	14
Tiered Access	15
Automation.....	16
Training and Credentialing	16
State, Local, and Tribal Issues	16
Core Functionality	17
Part 5 – Ideas for data concierge services.....	18
Data discovery services	19
Centralized assistance for data access	19
Chatbot for general inquiries	19
Data access services	19
Centralized assistance for navigating legal requirements	19
Anonymized queries on restricted data.....	19
Data use services	19
Statistical expert consultations	19
Library of data use best practices.....	19
Part 6 - Findings by persona	20
Persona 1: Resolve a Question.....	21
Persona 2: Access Existing Data	22
Persona 3: Create a New Linked Data Asset	24

Part 7 – Current and future service models	25
Service models – current state	25
Service models – future state	26
Part 8 – Review of existing data concierge services	27
Data infrastructure providers	27
European providers	27
Domestic U.S. providers	28
Appendix	30
Outreach Materials and Interview Guides	30
Federal Statistical Agencies	30
Data Users	31

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-DRN25-003.

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) has contracted with NORC at the University of Chicago (NORC) to explore models for a data concierge service for a potential, future National Secure Data Service (NSDS). The purpose of this study is to recommend approaches for providing a range of high-quality services to support data users who are pursuing evidence building research. This Environmental Scan and Outreach Report presents the detailed findings from our in-depth interviews with federal statistical agencies and non-federal data users, as well as our review of data concierge services provided by data infrastructure providers. This report serves as the basis for development of our next deliverable, the Data Concierge Model report.

Introduction

This report is organized around the data collected throughout the environmental scan and outreach task. Parts 1 and 2 of the report provide detailed information gathered from the federal statistical agency and data user interviews. We present our findings in Parts 3, 4, and 5, including the challenges facing data users, opportunities to expand user support, and ideas for future concierge services. The findings are organized around the essential elements of data use – data discovery, data access, and data usage, and their subcomponents. Part 6 illustrates the broad range of data user needs – through generalized personas – and how a future concierge service could meet those needs. An overview of the current service model and what a future service model may look like is provided in Part 7. Finally, Part 8 outlines our review of current data concierge services, both domestic and international.

Interview Methodology

NORC conducted 26 total interviews with respondents from federal statistical agencies (9 interviews) and non-federal data users (17 interviews).

Federal Statistical Agencies

NORC conducted 60-minute interviews with 9 federal statistical agencies (Bureau of Economic Analysis; Bureau of Justice Statistics; Social Security Administration; Census Bureau; Internal Revenue Service; National Center for Health Statistics; Bureau of Transportation Statistics; Energy Information Administration; and National Center for Science and Engineering Statistics). Of the remaining 4 agencies, two were excluded from participating and two others did not respond to our request. The interviews were conducted between October 30, 2023, and April 22, 2024. NORC and NCSES

collaborated on identifying ideal candidates to ensure each agency was well represented and the information provided was most useful. The goal of the statistical agency interviews was to determine what, if any, support services are offered to data users and to identify gaps a concierge service could help fill.

NORC initiated outreach by email and followed up with any non-responders, also by email, to schedule the virtual interviews. NORC sent an additional email to all non-responders. For any agency who did not respond after the third email, NORC identified an alternate contact and conducted the outreach protocol described above.

NORC conducted semi-structured interviews using a protocol that was developed in collaboration with NCSES. The interviews were conducted via Zoom with a primary interviewer leading the questioning and the secondary interviewer taking notes and asking follow-up questions as needed. The protocol (see Appendix) collected information on each respondent's role and their agency's approach to data access, any specialized assistance provided to data users, whether metadata is publicly available, the agency's ability to respond to user requests related to the Standard Application Process (SAP), staffing and resource allocations, and opinions about future services and tools that could help agencies better respond to data user requests.

Non-Federal Data Users

NORC received OMB approval to conduct interviews with non-federal data users on January 19, 2024. We conducted the 90-minute interviews with a broad array of data users from traditional and non-traditional research communities. In collaboration with NCSES, 27 organizations and individuals were identified within four broad categories: research/non-profit organizations, community advocacy organizations, state/local governments, and economic development organizations.

The data user interviews were conducted between March 7, 2024, and May 13, 2024. The goal of the interviews was to provide additional context regarding the availability of data related federal support services, which included identifying challenges and limitations to the provision of those services and understanding what types of services, support, and information would be most useful to data users. Before beginning outreach, NORC and NCSES organized the 27 organizations into three priority groups by relative importance. Outreach to Priority Group 1 (12 organizations) began in February 2024. After completing 10 interviews with this group, we began outreach to Priority Group 2 (10 organizations) and completed an additional 7 interviews. Given the high participation rate among the first two groups, we did not need to reach out to Priority Group 3.

	Research/Non-Profit Organizations	Community Advocacy Organizations	State/Local Government	Economic Development Organizations
Organizations Identified	14 organizations	3 organizations	8 governmental organizations	2 organizations
Interviews Completed	7 completed	2 completed	6 completed	2 completed

NCSES initiated outreach by email and NORC followed up, also by email, to schedule the virtual interviews. NORC sent one additional email to those who did not respond to the first two outreach attempts.

NORC conducted semi-structured interviews using a protocol that was developed in collaboration with NCSES and approved by OMB. The interviews were conducted via Zoom with a primary interviewer leading the questioning and the secondary interviewer taking notes and asking follow-up questions as needed. Prior to the interview, respondents were sent a list of questions as well as the participant consent information form. Before beginning the interview, all respondents were read the OMB-approved introduction and were asked to provide their consent to participate. The protocol (see Appendix) collected information about the respondent’s role and use of federal data, search strategies for identifying federal data (including awareness of the SAP and SAP Portal), data discovery resources, experiences with data concierge-type support services, and reactions to potential data concierge models, including specific services and tools.

Executive Summary

The goal of this report is to capture information about federal support services and data concierge services offered both within and outside the federal government, to identify challenges and limitations (because of policy requirements, bureaucratic practices, or other reasons) to the provision of these services, and to learn what types of services, support, and information is most important to a future data concierge service. To achieve this goal, we interviewed representatives from federal statistical agencies to determine the existing data user support services offered and elicit input on gaps and improvements to those services. Next, we interviewed current and potential federal data users to learn about the support services available to them in the discovery, access, and use of federal data. We also solicited their input on specific services and tools that could improve the process and expand the uses of federal data for evidence-based projects. The information collected for this report will inform the next phase of the project, i.e., the development of the proposed data concierge models.

The content of this report includes our findings on the current state of federal support services for the use of federal data and potential areas of where a data concierge service could improve and expand

those services. Our high-level findings on the current state of federal data user support services include:

- Data services provided by the statistical agencies are not standardized and vary in scale and scope.
- Statistical agencies and data users acknowledge that legal and policy restrictions on many sources of statistical data will continue to limit accessibility for evidence-based research. Efforts to improve access, such as a tiered access system within a potential NSDS, should be explored to mitigate those restrictions and expedite the evidence building process.
- Data users, especially those at the state and tribal level, acknowledge the potential value of federal data. Unfortunately, they are generally not resourced to navigate the process to discover and access those data until more repeatable and standardized processes are put into place.
- Experienced data users find that established source agency contacts are one of the best ways to discover relevant data and to gain assistance in gaining access to those data. Less experienced data users tend to rely on legacy data assets and find it difficult to identify and access new data sets.
- Data concierge service models should optimize automation to serve the full range of data users (from new users to power users) in the discovery and the path to access data.

Part 1 – Common themes in federal statistical agencies’ support for their data users

This section summarizes data discovery, access, and use considerations informed by interviews with federal statistical agencies listed below. During the federal statistical agency interviews, we examined the scale and scope of services they make available to potential users of their agency data. We compared the current services against data user requirements (derived from our environment scan) of data users to determine the “as-is” offerings of NSDS-like services available and the gaps that a data concierge service might fill. Those proposed services are contained in [Part 4](#) and [Part 5](#) of this report.

- Bureau of Economic Analysis (BEA);
- Bureau of Justice Statistics (BJS);
- Social Security Administration (SSA);
- Census Bureau;
- Internal Revenue Service (IRS);
- National Center for Health Statistics (NCHS);
- Bureau of Transportation Statistics (BTS);
- Energy Information Administration (EIA); and
- National Center for Science and Engineering Statistics (NCSES)

Data Discovery

Determining fitness for use

It is uncommon for agencies to directly support users in determining data fitness for use. However, agencies often provide some level of user support, which varies across agencies. For public data, agencies tend to provide documentation on their website and answer users' emails or calls. For restricted data, agencies told us that an application process is used to determine whether the data are well-matched to the user's request and often, to help narrow the variables that users are requesting. Agencies described how researchers who identify data in papers will reach out to authors for guidance before approaching agencies. Agencies with dedicated staff who have available time will iterate on research questions with users and help connect them to appropriate resources. Within larger agencies, there is so much data that it is hard for one person to have exhaustive knowledge of all assets, so support staffing tends to be more distributed. Agency staff noted that user requests can become complex when the requested resources exist across agencies. Sufficiently-resourced agencies said they help users identify the correct agency if their agency does not collect relevant data. Generally, agencies want their staff to have wide-ranging knowledge encompassing data across the agency and extending to relevant data offered by other agencies to help direct users. Staffing appears to shape agencies' ability to offer support and determines how interactive that support can be.

Availability of metadata

Agencies described providing a range of metadata and cataloging services. Some agencies have dedicated staff such as data curators, cataloguers, and metadata librarians who maintain agency data assets. Some agencies expressed frustration with gaps in their metadata; for example, it can be difficult to document known limitations of a data asset and recommend appropriate use (e.g., based on level of measurement). Further, agencies lack mechanisms to inventory how data have been used in the past (e.g., links to relevant publications). Some agencies are making metadata in the SAP Portal more detailed to support data discovery and proper use; however, this treatment of metadata is not standard across agencies.

On-line inventory/portal

Nearly every agency described their public website or Data.gov as the most complete and up to date inventory of their agency's public data assets. Some agencies described recent efforts to revamp their public websites to facilitate discovery, navigation, and ease of access (e.g., by listing distributed resources in a single centralized clearinghouse). Several agencies also referenced the SAP Portal as containing their most complete inventory of metadata describing all available statistical data assets. While some agencies also maintain separate inventories, the SAP was often described as the most comprehensive resource listing both public and restricted assets. Keeping metadata in the SAP Portal

current and synchronized as data assets evolve was mentioned as a challenge. This is currently a manual process. Some agencies also wanted ways to customize or enhance metadata in the SAP to better align with their own standards. Agencies also want a consistent look and feel of data assets across agencies and point to a need to first adopt shared metadata standards and protocols to enable this.

Data Access

Data accessibility

Most of the feedback on data access pertained to restricted data. The only concerns raised about public data access related to the means of data access. Modalities for public data access included file downloads, APIs, online analysis, and reporting dashboards; restricted data access included secure download, virtual computing environments, enclaves, and custom analysis platforms. Some agencies offer more modes of access than others and speculated that this might broaden their user base. For restricted data access, many agencies rely on the SAP Portal to guide users and explain criteria for accepting or rejecting applications. Some agencies find it difficult to manage the SAP and rely on a partner entity to track and review restricted data applications in the SAP Portal. Agencies are required by law to build a “revise and resubmit” step into the SAP to guide users through obstacles encountered in the process, which reduces the total number of rejected or failed applications. However, despite having a “revise and resubmit” process, agencies vary in the extent of support they provide to applicants, with some engaging in more intensive assistance (e.g., in crafting research questions and design) and others providing less intensive support. The differences appear largely to be driven by staffing considerations. Agencies also provide dedicated staff as a point of contact to explain access criteria and help researchers meet those criteria.

Gaining access

Agencies agreed that researchers from statistical agencies and large research institutions tend to be more experienced in gaining access to restricted data. Agencies mentioned that strategies, such as having a federal partnership or collaboration, can also help users negotiate access, and that researchers from R1 universities tend to leverage this strategy to gain access to restricted data. Agencies expressed concerns about a lack of equity in gaining access to restricted data across institutions of higher education as well as state and local governments. In addition to enabling continued access to data by other federal agencies and experienced researchers at R1 universities, agencies expressed a desire to create additional pathways for new users from state agencies, smaller non-R1 institutions, and nonprofit organizations to gain access to restricted data. Data can be made available through a secure download or an enclave (e.g., virtual, remote). Remote access to restricted data is increasing, but some agencies have limits, such as on the number of “seats” available per terminal and associated service fees, that limit remote access to restricted data in a virtual enclave. Further, even if researchers qualify for access, it may not be granted, or there may be delays because

of the overwhelming number of applications that must be reviewed or the intensity of requests for support that come with those applications, such as requiring a curated dataset.

Securing legal/policy authority to access

While some agencies still consider scientific merit when reviewing requests for data access, others have dropped this requirement, for example noting that scientific merit can be difficult to interpret and evaluate across various disciplines. As justification, the Evidence Act and the presumption of accessibility clause of CIPSEA offer provisions that are equal to or greater than those of scientific merit review. As additional considerations that govern data access, some agencies also mentioned compliance with specific titles, MOUs with states, state laws, and even FOIA requests as a form of specialized assistance for potential users.

Data Usage

Authority to publish

Agencies that publish data products described that the products should support many use cases. Some agencies with limited authority to publish restrict access by sharing their data with organizations instead of individuals. Other strategies include pairing staff, who review projects for disclosure risk, with researchers and curating custom datasets and tabulations to support specific research projects. Staff involvement helps to ensure that data are being used appropriately and that disclosure risk is minimized throughout the project. Examples of data that are considered too sensitive to share include administrative data from within agencies and microdata. Agencies often need to manipulate individual-level data prior to publication or sharing, but there are concerns that aggregation, masking, or other procedures to remove or alter personal or business identifiable information (PII/BII) and geographic identifiers may limit the research utility of data products. Agencies referenced tiered access as a strategy that could help to determine the kind of data a user might need and match requests to their appropriate level of access along a continuum, from synthetic derivatives to individual level records.

Ability to link and link keys available

Some agencies mentioned that they wanted to build out their PPRL (Privacy Preserving Record Linkage) capacities and upskill staff to build personnel expertise in data confidentiality and disclosure risk. Many agencies seemed aware of the potential to enable more record linkage but were only able to provide limited support due either to privacy concerns or staffing constraints. Most of the support for linked data is provided in the form of existing agreements between agencies to link data assets (e.g., between the BEA and Census). In certain cases, records include alternate keys, such as a protected identification keys (PIKs), that allow records with sensitive PII to be linked to other internal databases. However, Title 13 requirements prevent PIKs from being introduced outside of the Census/FSRDC

domain. Record linkage is also enabled through secure access mechanisms, such as networked servers, where users can analyze linked data.

Determining quality of linkage

Agencies mentioned the importance of subject matter expertise to determine the quality of links. Agencies are also more familiar with their own data assets and are less able to evaluate linkage quality for external data. Quality was often described alongside fitness for use and determined by use case; a level of quality may be sufficient for some applications but inappropriate for others. Record linkage is sometimes performed in-house by staff on a case-by-case basis, which provides quality assurance for the final product. Agencies that offer comprehensive, coordinated linkage services on a case-by-case basis have a limited capacity to provide these services, which are not scalable.

Part 2 – Common themes in federal data users’ experiences with statistical data

After determining the services statistical agencies make available to their users, we conducted a different interview with current and potential users of federal data to compare their experiences and requirements against the current federal data support services ecosystem. This section summarizes data discovery, access, and use considerations within the current “as is” infrastructure. These results are informed by interviews with seventeen federal data user users from the following organizations:

- State Health Access Data Assistance Center (SHADAC) at the University of Minnesota;
- Federal-State Cooperative for Population Estimates (FSCPE) – South Carolina, New York, Colorado, Washington;
- National Archive of Criminal Justice Data (NACJD);
- Inter-university Consortium for Political and Social Research (ICPSR);
- Chapin Hall at the University of Chicago;
- Federal Statistical Research Data Center (FSRDC) users (2);
- Integrated Public Use Microdata Series (IPUMS);
- National Association of County and City Health Officials (NACCHO);
- National Congress of American Indians (NCAI);
- Massive Data Institute (MDI) at Georgetown University;
- National Institute of Standards and Technology (NIST);
- Economic Development Administration (EDA); and
- Actionable Intelligence for Social Policy (AISP) at the University of Pennsylvania

As-is Data Discovery

Determining fitness for use

Data users cited data accessibility, availability of data documentation, file formats, user support, and quality issues as factors that help them determine data fitness for use. Users who work with data from multiple federal agencies expect the reliability, accessibility, and level of user assistance to vary. When navigating public use data, many users find agency websites difficult to use. In general, users expressed a need for more context, such as explanations of which data are currently available and why. Determining fitness for use of restricted data can also be challenging. Many users are unwilling to spend time trying to get access to data without knowing if they will be allowed to produce an answer to their question or publish the result (e.g., due to small cell sizes in resulting tabulations). File formats can also deter users from working with specific datasets given the level of effort that would be needed to extract, transform, and load relevant information.

Users have different strategies for getting answers to their questions about data. Users often reach out to designated agency staff, to their peers, or rely on existing data-related literature. When users approach agency staff, reference interviews (i.e., conversations between staff, such as librarians, and a data user through a reference desk) can help users express their needs and adjust their questions and requirements. However, users expressed frustration when agency staff were not deeply familiar with the data and were unable to answer users' questions. Researchers did not find general inquiry forms helpful, especially if they had expertise on the topic. Users mentioned that the timeliness of responses to user inquiries is also critical in informing whether they will pursue data. If users cannot get an answer in a reasonable amount of time, they may decide to go a different route even if the data would be fit for purpose. Researchers said that they were often better served by talking with peers who have previously worked with the data for guidance on use. In addition, users benefit from seeing which papers have used a particular dataset or examples of datasets that have been used together helps researchers. Users sometimes resort to "reverse engineering" solutions by reviewing published papers that use the data to see how to work with it.

One user described their worst-case scenario as being able to get access to data but being unable to get answers to questions about how to properly interpret or use the data. This is especially concerning when users uncover quality issues with data. Some users felt that data quality issues have become more apparent since the pandemic; for example, they have encountered federal data that are out of sync or have incomplete or uninterpretable documentation. Users often do not discover these limitations until they are well into their analysis and need guidance in the form of data documentation or an agency contact, which is not always readily available.

Determining availability of metadata

Many data users agreed that, at a minimum, all available data should include a codebook, a data dictionary, or a read me file. Users described that working with data assets that lack proper documentation slows down their analysis. Metadata enables users to decide whether to invest time in working with data or move forward with pursuing access to a particular data asset. Metadata help users search for data and determine data fitness for use. Granular information, such as units of observation and levels that estimates can support (e.g., for a power analysis), and high-level information, such as data dictionaries and record layout, are often needed to help users make this determination. Users want to know how data were collected, their purpose, and who was responsible for data collection, especially for administrative data. Users agreed that there should also be documentation of known issues or limitations of data. In addition, users want clear explanations about which variables are suppressed and why included in the metadata make it easier for researchers to appraise data. Several users mentioned that metadata with contradictory or misleading information, such as variable names that do not match the dataset, can be worse than providing no documentation at all.

Reviewing online inventories

Generally, users described dataset searching as time intensive. Researchers tend to start from literature reviews in their topic of interest to identify unfamiliar datasets and understand their reuse potential. Others take a “bureau by bureau” approach, employing different strategies for each agency’s site based on their level of familiarity with the agency. Keyword-based searching on agency websites has limited utility. Users supplement their searches by targeting topical (e.g., ICPSR) and/or foundation data repositories (e.g., Kaiser Family Foundation). Sometimes this approach can be productive. One user described the experience of “stumbling” into relevant data from an unfamiliar agency by browsing for their research topic in an aggregator website (e.g., Centers for Medicare & Medicaid Services ; Research Data Assistance Center). Many users tend to rely on Google searches as a starting point instead; however, some expressed concerns that this strategy tends to return large data and does not usually surface smaller, less frequently used data. When all else fails, users turn to their social networks for guidance. Data “power users” turn to their communities of practice, which specialize in particular domains but have incomplete knowledge of existing data. Many users agreed that word of mouth and “knowing the right people” is often key to identifying data; however, keeping track of who to contact is difficult when staff in agencies leave or change positions.

As-is Data Access

Experiences of data accessibility

Access to data is often a necessary precursor to determining fitness for use and evaluating quality. Users described barriers to data accessibility including agency staff who provide incorrect information about data assets, publication formats that are not machine actionable (e.g., hundred-page PDFs with

tables), and missing documentation. The size of some data assets, especially those provided through secure enclaves, and the required analytical tools (e.g., opening spreadsheets with millions of rows in Excel) can make data use and analysis prohibitively time consuming and frustrating. When data are made accessible without documentation or knowledgeable support staff who can answer questions, users interpret this to mean that the agency providing the data does not care if the data are used or not. Some users perceived this as obfuscation and mentioned that it also made them feel less confident in the data quality. When exhaustive searching is not possible, users may satisfice by selecting datasets that are easier to work with and are readily accessible.

Gaining access

Most data users viewed gaining and providing access to federal data as central to their role; some also viewed their organizations as data concierge services that provided technical assistance, for example by using federal data to answer questions for state policy makers or federal employees. In this capacity, “data scans” are an important part of many data users’ jobs. They must maintain a high level of awareness of what data are available in their domain and be knowledgeable about how to access those data, what the limitations of the data are, and make determinations about data quality to recommend their use. Some users spend a great deal of effort trying to get access to data. Many users described knowing that data exist but not knowing how to get access to them, including a lack of familiarity with the SAP. In this regard, the SAP appeared to be underutilized.

Securing legal/policy authority to access

Some of the organizations that data users represent help customers navigate data use agreements when working on projects with federal data. Users mentioned a greater need for disclosure guidance, particularly at the outset of acquiring data, and agreed that standardized services that provide guidance across agencies would be valuable.

As-is Data Usage

Authority to release tabular results

Data users’ reactions to services that would produce and release tabular results were mixed. In the absence of specialized tools or services for working with data tables, navigating documentation, and searching through tables published on federal data websites can be time consuming and complex. Users struggle to work with data tables published in PDF format, which creates obstacles to accessing and analyzing data. Some users described leveraging existing tools, like the Federal Reserve’s FRED platform¹, to generate high level tables and figures from data based on queries. Many felt custom table outputs produced upon request would be a useful service. This need could be met either through

¹ <https://fred.stlouisfed.org/>

secure multi-party computing (i.e., where a user is able to produce their own custom extracts from secure data) or through requests to staff for custom tabulations. Users did not have experience with producing their own custom extracts using secure multi-party computing, but most users had either worked with agency staff to request custom tabular extracts or had produced their own extracts after gaining access to restricted data. Tabular outputs are also used to quickly compare data sources (e.g., for validation). However, some users saw the production and release of tabular results for customers as central to their roles and viewed a federal service that produces tabulations as a competitor that might “put them out of business.”

Authority to link and link keys available

Several users we interviewed work for organizations that perform linkage on other users’ behalf, especially for restricted data. They described that some reliable federal data sources for linking take a very long time to gain approval. Even after approval, record linkage can take a long time because permitted linkages need to be spelled out and results need to be submitted for disclosure review. Projects that use federally funded data tend to limit or restrict linkages. Another common scenario is linking on a standard geography, but users expressed that certain methods such as differential privacy adjustments limit the ability to standardize geographic data.

Determining quality of linkage

Quality indicators for linked data are often lacking and are not easily determined without access to the original data. Users described that they find it hard to understand or evaluate the success rate of linkage strategies. Users expect match rates to be well-documented and shared. When using linked data, users want to understand how data were processed and which linkage methods were used to make determinations about the reliability of the data and its fitness for their particular purpose.

Part 3 – Challenges faced by data users

This section summarizes data discovery, usage, and access challenges uncovered in the federal statistical interviews ([Part 1](#)) and data user ([Part 2](#)) interviews. The quotations included below are illustrative and are intended to help paraphrase multiple pieces of feedback from interviews.

Data discovery challenges

Searching for data

- “I have a hard time searching for data across agencies by topics like employment.”
- “I know what variables I’m interested in, but I’m not sure how [to] search for them in an agency.”

- “I don’t know what I don’t know, so I feel like my searches are incomplete, or that I’m either looking in the wrong place or looking for the wrong thing.”
- “I know the sources that I’ve worked with well, but I have a hard time looking for new data.”
- “I don’t have the time or staff resources to dedicate to searching for novel data.”
- “I tend to look for data from literature so I can ‘reverse engineer’ my search by seeing what topics and analysis the dataset supported.”

Determining data fitness

- “I often struggle to understand what the data contain or how they can be used when metadata aren’t detailed enough... I usually have to download the data and explore them.”
- “I can’t understand data without descriptive codebooks, variable lists, or other documentation.”
- “Outdated, incomplete, or contradictory metadata gives me pause... I either need to reach out to the data provider to ask for more information, or I might decide not [to] use the data at all.”

Data access challenges

Navigating restricted access

- “Without an agency partner, it can be difficult to qualify for access to restricted data or to know what steps to take to qualify for access since the process can vary by agency.”
- “Users requesting restricted data who don’t meet requirements must go through a ‘revise and resubmit’ process to modify their request”, which delays the data acquisition process.
- “Requesting restricted data from different agencies can be repetitive and time consuming.”

Meeting training and credentialing requirements

- “CIPSEA requirements could open presumption of accessibility by authorizing data use more broadly but are currently used by many agencies to restrict data access.”
- “Few agency staff have extensive confidentiality and disclosure risk (Title 3) training, placing burdens on trained staff to help users with restricted data access needs.”

Data use challenges

Validating statistical analyses

- “I find myself wondering if others have done a similar analysis with this data and whether I’m on the right track.”

- “Sometimes I find that I have questions about data that I discover during my analysis but I’m not sure who to turn to.”

Linking data

- “I understand the potential for linking data, but I’m not sure which data are eligible or how to go about setting up a project that does this.”
- “When I work with linked data, I find it hard to evaluate the quality of the linkages for my needs.”
- “I’m interested in PPRL but the agency I’m working with either doesn’t allow or support it.”

Reviewing disclosure risks

- “Disclosure requirements were not available prior to beginning analytical work, and after review I’m unable to release the findings as I had planned.”

Part 4 – Opportunities to support data users

Our interviews with federal statistical agencies and data users revealed opportunities that could help mitigate the challenges that users face ([Part 3](#)). The features for a data concierge and/or a future NSDS more broadly described in this section encompass such opportunities.

Data Discovery

- **Enhanced search:** Allow users to search for data across agency sites, Data.gov, and the SAP Portal.
- **Federated catalog:** Provide a complete, up-to-date inventory of agency data assets.
 - Communicate inventory “completeness” by participating agency (e.g., leaderboard).
- **Variable-level and codebook indexing:** Enable users to search for granular variables.
 - Support the comparison of variables across datasets and include definitions.
- **Linked code and documentation:** Improve statistical agency capacity to retain analytical code and associate data with user documentation to support reuse.
 - Include codebooks with survey questions, survey instruments, lists of variables, and “0 observation” (empty datasets).
- **Custom metadata:** Enable additional metadata fields in the SAP portal so agencies can more closely replicate their existing metadata.
 - Communicate what the data measure and how they should/shouldn’t be used.
- **Data “gems”:** Create video guides highlighting case studies with particular data assets.

- **User forum:** Provide a site for users to pose, answer, and rate data-related questions.

Tiered Access

- **Data linkage:** Develop a process to help users understand and evaluate the quality of data linkages.
 - Perform record linkage on behalf of the user.
 - Provide an environment for data linkage with IT guardrails against nonsensical linkages.
 - Inventory eligible datasets, explain how to use them, and how to apply for access.
 - Expose standard link keys so users don't have to reinvent the wheel.
 - Help users understand and evaluate quality of data linkages.
 - Create tutorials or guides walking through record linkage vignettes.
- **PPRL:** Develop a privacy preserving record linkage process approved by statistical agencies.
 - Perform PPRL on behalf of the user.
 - Integrate a vendor solution that enables users to perform PPRL in a secure environment.
- **Disclosure review and mitigation:** Enable users to request tabular outputs from restricted data.
 - Create custom extracts from restricted data based on user specifications.
 - Add an interaction controller to accept and refine user queries, pull, and clear data based on the query, and provide sanitized data back to user.
- **Anonymized queries:** Provide information about data without providing direct access to data.
 - Offer standardized queries with minimal disclosure that satisfy users' requests.
 - Provide a query library with possible analysis workflows pre-loaded.
- **Wiki:** Enable editing that allows users to update (meta)data and leave feedback.
 - Capture the adjustments that researchers make to data that are not preserved beyond the original analysis.
 - Curate entries to mitigate confidentiality risks for person-level data.
- **Invoke/apply CIPSEA requirements:** Open presumption of accessibility by authorizing data use more broadly.
- **Synthetic data repository:** Create synthetic versions of files that are analytically useful.
 - Inventory all available synthetic datasets and their fitness for use.
- **Data acquisition templates:** Offer boilerplate language and a stepwise DUA process.
 - Integrate the process with the SAP.

Automation

- **Recommend access level:** Interpret user requests and recommend which level of tiered access they would need and why.
- **Metadata synchronization:** Track the lineage of different data sources and how metadata have been updated between agencies repositories and the SAP Portal.
- **Sample checks:** Run and report results of tests on restricted data.
- **Build bibliographies:** Identify dataset citations from research articles.
- **Extract, transform, load (ETL) service:** Dynamically extract source data from a source system, transform it, and offer it in a user-friendly format (tables, shapefiles) based on users' needs.
- **Bots/agents/chatbots:** Mine text and produce relevant crosstabulations.
 - Host within restricted enclaves to provide alternatives to web-based search in restricted computing environments.
 - Replicate a “reference interview” structure (between a researcher and librarian) to elicit requirements and needs from the prospective data user.
 - Generate graphics based on a question like the Federal Reserve’s FRED tool.
 - Learn lessons from libraries and states who experienced staff reductions and used digital tools to replace those previously performed by humans.
 - Improve trust in the Federal Statistical System by improving the quality of interactions and information obtained.
 - Rate requests as “easy”, “medium”, or “hard” and triage accordingly.
 - Use expert “humans-in-the-loop” to assess quality and improve responses.
 - Provide 100% factual responses sourced from up-to-date documentation.

Training and Credentialing

- Standard, portable CIPSEA training for qualified users.
- Confidentiality and disclosure risk training for staff (Title 3 requirements).
- **Researcher passport** for qualified users who have completed a background check and training.
- **Audit trails** about who has used data and how to share back with data providers, particularly to promote transparency around tribal data use.

State, Local, and Tribal Issues

- **Expand access** to state agencies, small institutions (non-R1), nonprofits.

- Address concerns about broadening the user base and reduce the “clubbiness” of the data reuse community. Experienced data users, such as academics from well-resourced research universities, tend to capitalize on existing research relationships with federal agencies and face fewer barriers to accessing and using federal data, especially restricted use assets. Establishing clear, standard pathways and access requirements for all users, regardless of their affiliations or existing collaboration networks, to apply for access to restricted use data may help shift perceptions and encourage more users outside of R1 universities to seek access to federal data.
- **Build trust** by including explanations of how data have been produced and should be used.
 - States, local governments, and tribes in particular, expressed how a lack of transparency in how data are distributed and reused reduces participation and trust in data collection processes. Adding guardrails around data sharing and including explanations of how data should be used can help enhance trust.
- **Share services** with state, local, and tribal governments, which often do not have the same capacity to deploy tools and services.
 - Create modular, templated solutions with shared ownership.

Core Functionality

- Support to data users should be universal (to users), standardized (across agencies), and repeatable (process).
- Better integrate data users with agency “learning agendas” to improve data access.
- **Connect users to experts:** Users should be able to come up with a research question and get connected with someone who has research expertise in that area.
 - “Connect existing dots” of people and places with expertise.
 - Understand the “universe of what’s available” and help a user determine whether an answer could be found in published data.
 - Help new staff “learn the ropes” and build institutional knowledge.
 - Guide users in the right direction and direct requests to relevant experts to save time.
 - Respond in a timely and tailored manner for best customer experience.
- **Integrate with existing tools:** Services should be well-integrated into the communities that they serve by interfacing with familiar platforms and meeting users “where they are.”
 - General audiences tend to use Google search, Data.gov.
 - Research audiences use word of mouth and are starting to use the SAP, though there is a need to increase the visibility of the SAP outside of power user communities.
- **Support data evaluation:** Help users evaluate different data products by showing additional context information, such as how data were produced and for what purpose.

- Help refine research projects and streamline access to data by helping users identify appropriate datasets and processes for accessing them.
- Draw from grey literature, such as documents written for sponsors, that include lessons learned, failures, and findings that aren't otherwise published.
- **Serve customers:** Incorporate a customer relationship management (CRM) infrastructure.
 - Log events in a customer interaction database that can provide a source of training data and analytics for identifying gaps in services and assets.
- **Support cross-agency requests:** Support users by combining agency data.
 - Refer users to resources across agencies and help them way-find as they navigate from the SAP Portal into agency sites.
 - Submit a single request across multiple data sources.
 - Support researchers who draw from a wide variety of datasets (e.g., to study social and structural drivers of health outcomes).
 - Explain each agency's rules and disclosure authorities.
 - Walk a user through steps of applying for data across agencies.
 - Function like SpringShare > LibAnswers ("Ask A Librarian") chat service.
- **Communicate coverage and completeness:** A means of communicating how complete and current the inventory of data is and what plans exist to add additional sources
- **Create a corporate voice:** Effectively communicate and advocate service offerings
 - Provide a newsletter that highlights available services, communicates information on new improvements, lists contact information, etc.
 - Produce shareable documentation that removes "word of mouth" communication
 - Have a presence at relevant conferences, organization meetings, etc.
- **Provide governance:** An expert steering committee could guide services to meet user needs

Part 5 – Ideas for data concierge services

This section focuses on components that could inform a data concierge as part of a larger NSDS and is informed by feedback from our interviews with federal statistical agencies ([Part 1](#)) and data users ([Part 2](#)), as well as our scan of existing data concierge services ([Part 8](#)). We envision data concierge services interfacing with multiple parts of the NSDS, for example, by helping connect users to self-service and staffed services that support data access and use.

Data discovery services

Centralized assistance for data access

- Experts with general knowledge of the NSDS services and existing federal data assets
- Assist with data discovery process, focus on directing users to useful information
 - Utilize DCS tools to reduce resource requirements when interacting with users
- Provide direct assistance for unique requests by connecting users to agency staff or SME

Chatbot for general inquiries

- Interface to cross-agency FAQs and public-facing documentation
- Surface relevant agency contacts and related resources
- Help new users navigate across components of the NSDS

Data access services

Centralized assistance for navigating legal requirements

- Access to DCS staff with general knowledge of the legal requirements for data access
- Ability to interact with and refer users to agency staff for complex or challenging issues

Anonymized queries on restricted data

- Provide information about data without interacting with the data
- Standardized queries with minimal disclosure to satisfy users' requests
- Provide a query library with possible analysis workflows pre-loaded

Data use services

Statistical expert consultations

- Answer user questions that arise during analysis (e.g., statistical consultations)
- Connect users with subject matter experts to handle complex questions
- Provide information and discuss disclosure limitations at the beginning of the research process

Library of data use best practices

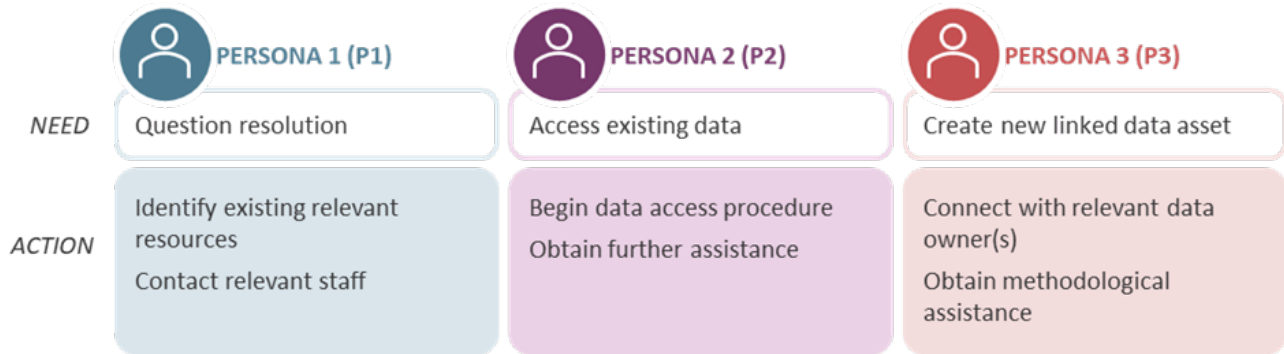
- Inventory of use cases and supporting documentation for data assets

- Links to video tutorials ("data gems"), data-related publications, grey literature, and code

Part 6 - Findings by persona

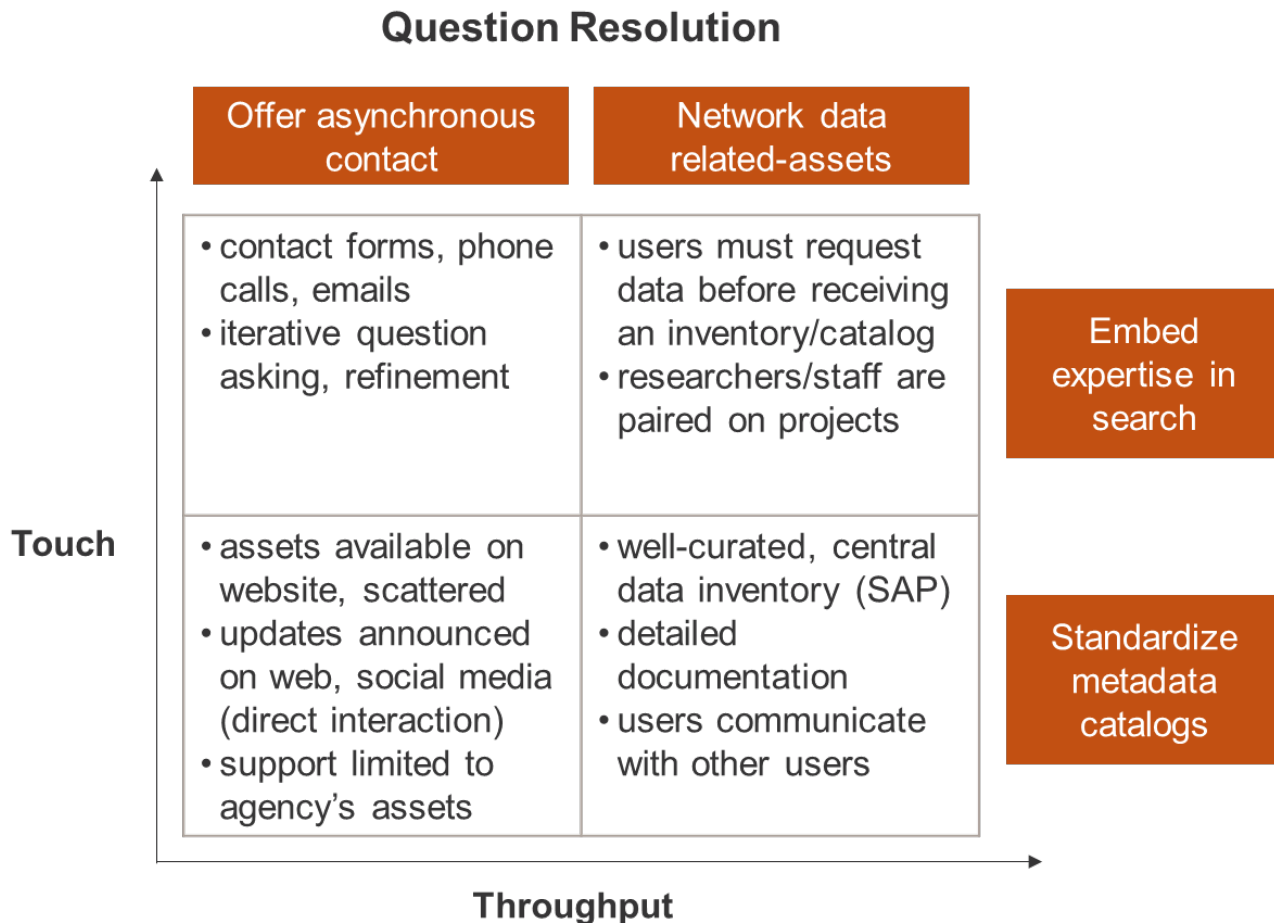
This section organizes findings from our interviews around three personas with different levels of interface or “touch” needs -- low is no direct interface; medium is some; and high involves direct interactions between users and staff -- and the volume of users these methods serve or “throughput” – low is either due to a limited or restricted user base or scope of service; medium has additional service capabilities but may be limited due to staffing or other constraints; and high means that many users engage with the agency’s data).

Exhibit 1: Visualization of ACDEB “generalized personas”



Persona 1: Resolve a Question

Figure 1. Requirements for identifying existing resources and contacting relevant staff

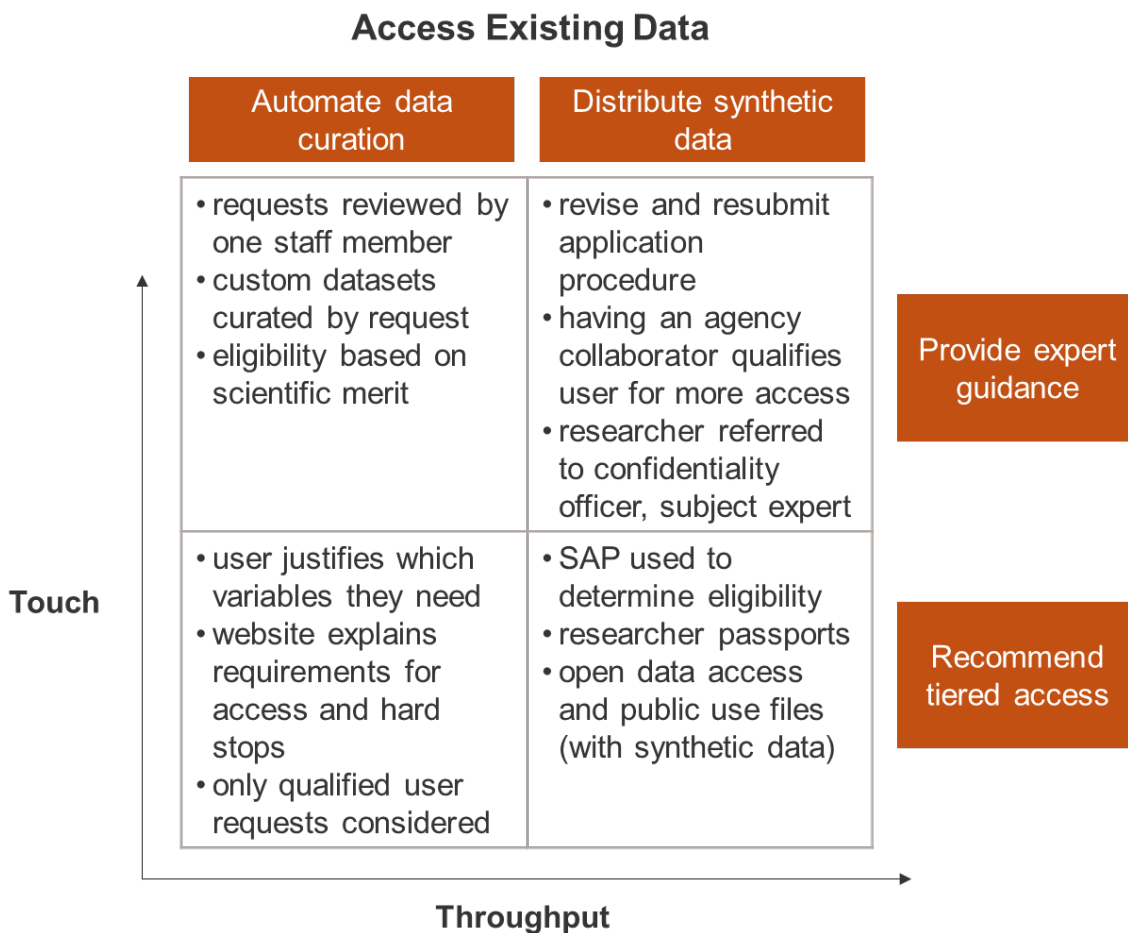


- **Low touch and low throughput** strategies for answering user questions include listing data assets on the agency website (often in a decentralized manner), posting announcements about data assets on the agency website or social media, and limiting question-answering to agency-specific assets (i.e., not responding to requests about assets outside of the agency). These approaches are low touch because they require minimal staff support. They are low throughput because they can limit potential data use (e.g., by restricting the scope of agency services).
- **High touch and low throughput** strategies include direct interactions with individual customers, such as researchers, through contact forms, phone calls, and emails. In these interactions, and when staffing allows, users can refine their questions as they learn. They are high touch because they involve one-on-one interactions between users and agency staff, and they are low throughput because they typically address questions on a case-by-case basis.

- **Low touch and high throughput** strategies include developing comprehensive inventories (such as the SAP portal) where users can review detailed documentation. While considerable investments must be made upfront to create high-quality metadata, procedures for maintaining detailed inventories can reduce demands on staff time by users over the long run.
- **High touch and high throughput** strategies involve either having staff work directly with users (embedded in research projects) or building capacity via networks of data users. In the first case, there is a high degree of touch between staff and users extending through the lifespan of the project; in the latter case, high touch exists between users themselves, such as researchers who have expertise in using specific statistical data assets. It may be valuable to surface data-related assets, such as research publications and authors, who share analytical details about data with other prospective users.

Persona 2: Access Existing Data

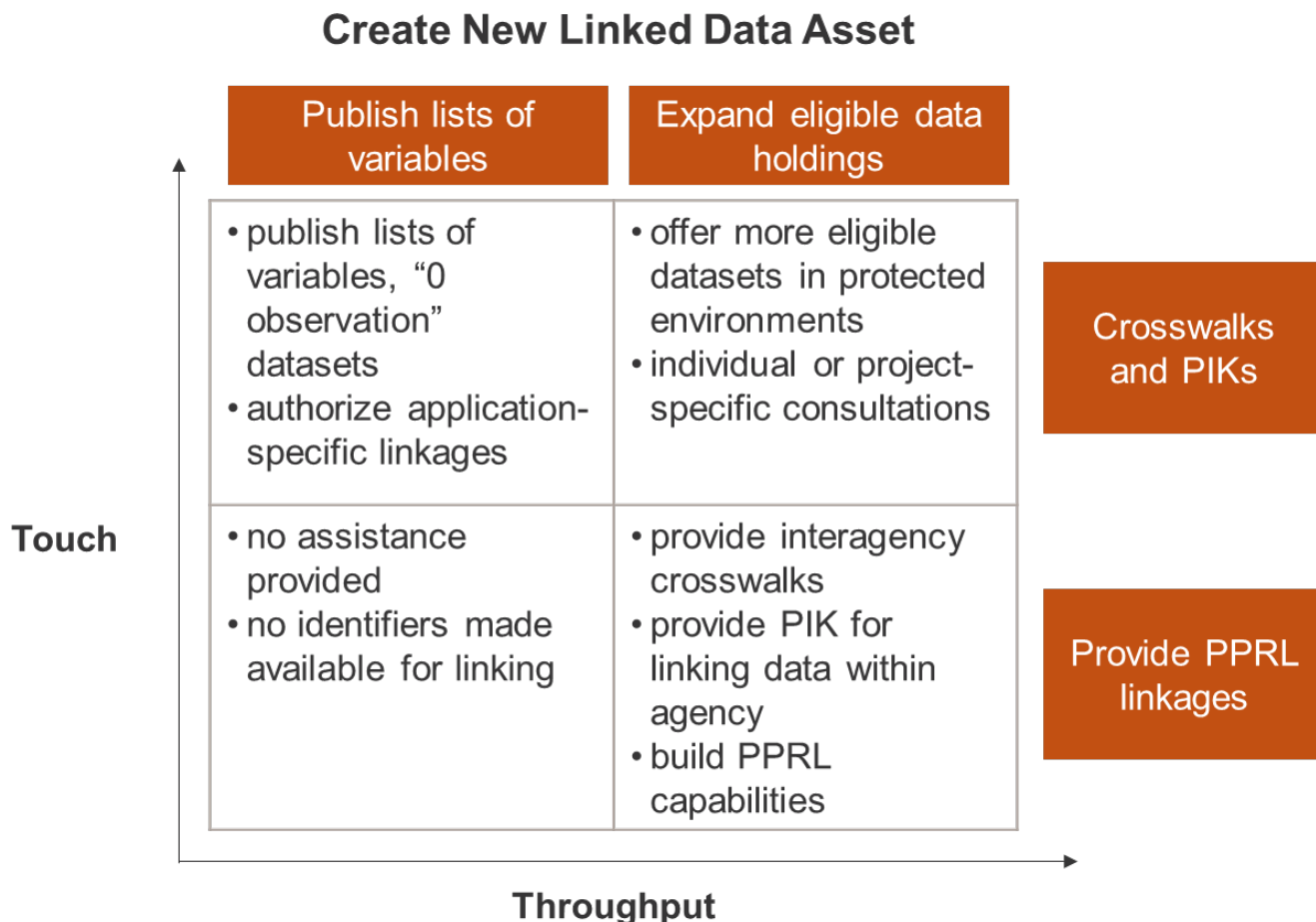
Figure 2. Requirements for beginning a data access procedure and obtaining further assistance



- **Low touch and low throughput** strategies for mediating data access include requiring users to explain which variables they need (assuming users have complete knowledge of available data), posting requirements on the agency website (rather than fielding them through dedicated staff), and only responding to qualified user requests. In this service model, the burden of understanding eligibility falls mainly on the prospective user. Making tiers of access more apparent in low touch scenarios may help users navigate eligibility requirements and understand the kinds of access that are available upfront.
- **High touch and low throughput** strategies tend to be handled by a single staff member and respond to requests for intensive efforts, such as the curation of custom datasets. Consequently, time-intensive requests can be subject to additional considerations, such as scientific merit, to justify the effort. However, merit can be difficult to determine in an interdisciplinary context. These scenarios are high touch because staff interface with users. Services that automate data curation (for example, by flagging and mediating PII in a dataset) in combination with user tiers could expedite data access for more users.
- **Low touch and high throughput** strategies rely on the SAP to determine eligibility requirements. Using the SAP to field applications reduces burden on agencies to provide specific guidance on eligibility requirements. Interview participants raised the idea of a “researcher passport,” which would allow qualifying applicants to share their credentials with other agencies and automatically demonstrate eligibility requirements. For open data, there may be initial costs associated with publishing but once available, the comparative burden and workload associated with providing access to the data is lower than for restricted data. Making more synthetic data available to satisfy particular use cases may be one way to provide low touch, high throughput restricted data access.
- **High touch and high throughput** strategies include revise and resubmit application policies, which allow prospective users to update their requests if they are flagged or denied by the agency. While this places a burden on the agency to review and respond to each qualifying application, it ensures that eligible users do not fall through the cracks and are able to modify their requests to gain access to needed data assets. Some agencies grant additional data access to users with agency collaborators. Users may also be referred to a confidentiality officer or subject expert who can assist with their request. Subject expert guidance could be provided following the models of agencies that currently employ dedicated subject experts to work with users, or could leverage emerging technologies, such as AI (Artificial Intelligence) chatbots, to provide expert guidance for inquiries.

Persona 3: Create a New Linked Data Asset

Figure 3. Requirements for connecting with data owners and obtaining methodological assistance



- **Low touch and low throughput** strategies include limiting or not offering assistance for data linking or record matching. This can be due to staffing limitations, lack of staff expertise with data privacy, or due to infrastructure limitations (i.e., not being able to provide identifiers for linking or a secure enclave where linking can take place). These strategies are low touch and low throughput because agencies do not provide staff who interface with users and do not make linked data available.
- **High touch and low throughput** strategies include publishing lists of variables available for linking and “zero observation” datasets so users can review data capabilities without accessing restricted variables. Application-specific authorizations can also be negotiated between agencies, paving the way for secure linkages that maintain data privacy. These strategies are low throughput because they either do not provide the actual data needed for linkage, or they authorize interagency data linkage on a case-by-case basis and do not make linked assets widely available to many users. They are high touch because they require a high degree of communication between agencies to negotiate agreements, or within agencies, to prepare high-quality documentation of variables.

- **Low touch and high throughput** strategies include providing interagency crosswalks, providing PIKs (Protected Identification Key) that allow record linkage within an agency, and providing capacity for PPRL, which would enable agencies to provide a range of linked data assets. While PPRL would require substantial effort upfront, it could offer a low touch and potentially high throughput avenue for providing access to linked data assets across agencies over time.
- **High touch and high throughput** strategies include providing staff available for project-specific consultations and offering guidance or services for data linkage. Offering more eligible products within data enclaves and secure computing environments would also create more possibilities enabling users to create new linked data assets.

Part 7 – Current and future service models

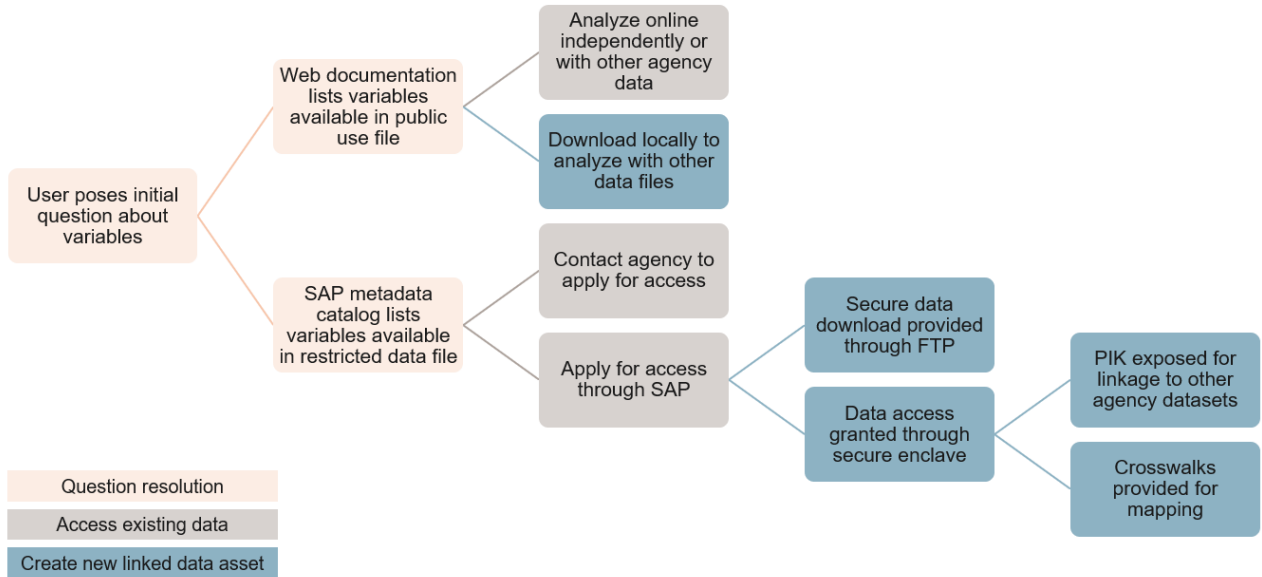
This section brings together the phases of question resolution, data access, and data linkage for comparison between current service models and future ones. We discuss interventions that were proposed during the interviews for each type of model.

Service models – current state

Figure 4 provides an overview of current state service models color-coded by user persona. The first steps in Figure 4 show users seeking to resolve a question. Some agencies have staff who can help direct user questions. Other agencies offer little to no direct contact with users; instead, users must find and review existing, publicly available documentation and data or metadata inventories to determine what they need. The more detailed the metadata and documentation are, the more likely users are to succeed in identifying appropriate data and their fitness for use. The next steps in Figure 4 show what happens once users determine whether publicly available or restricted data would be more suitable for their needs. Users are either able to download the data directly or analyze it online if the data are publicly available. If the data are restricted, users either need to contact the agency for or apply for access through the SAP and wait for a staff member to respond. Applications for restricted data may be initiated at any time, but some agencies only review applications at certain times of the year. The standard process guides users through several steps put in place by the agency to gain access to the data, pending approval. If granted access, an approved user would then be able to securely download the data or analyze the data in a designated environment, such as a secure data enclave where additional services, such as record linkage, may be enabled.

Figure 4. Alignment of personas with current state service models

Service models: current state

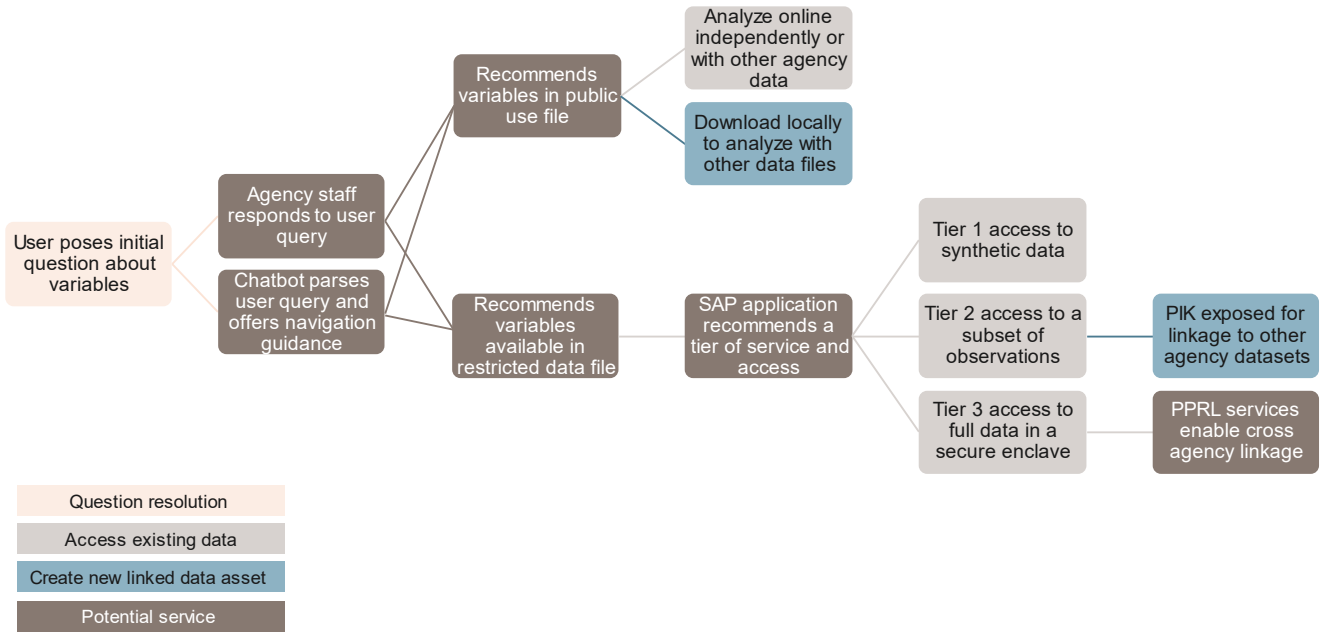


Service models – future state

Figure 5 provides an overview of several potential future state service models color-coded by user persona; the dark grey boxes reflect areas of opportunity where we can focus our efforts on developing the eventual proposed models. In the first step, a staffed help desk with agency contacts could serve as a first point of contact and guide users to appropriate resources or agency staff. Support staff could also help inform the next set of steps by interpreting users’ queries and determining whether the data that users need is publicly available or restricted. If restricted, then staff could provide consultation and guidance to walk users through the restricted data application process. Within the SAP, providing multiple tiers of access could allow more users to interact with restricted data assets. Ideas for what these tiers might look like range from synthetic data to full data access in a secure enclave where additional privacy preserving record linkage services might be offered. As part of this process, users might also be able to pose more complex questions spanning multiple agencies. Alternatively, a chatbot feature could provide timely intake services and help direct users to various resources on agency websites or quickly respond to users’ frequently asked questions. This service offering could potentially reduce staff resources necessary to respond to routine user inquiries.

Figure 5. Alignment of personas with future state service models

Service models: future state



Part 8 – Review of existing data concierge services

This section reviews concierge support services (both software- and expert-driven) offered by European and domestic data infrastructure providers on behalf of federal, state, and local agencies. NORC investigated these infrastructure providers because they often face the same or similar challenges that future data concierge service and NSDS would need to help resolve, and their solutions might provide novel insights into addressing challenges that federal agencies and data users currently face. We provide a brief description of each provider and highlight the types of solutions they provide to address challenges noted by interviewees. This information will also inform the proposed models we develop in the next report.

Data infrastructure providers

European providers

- EUDAT Collaborative Data Infrastructure: network of 25+ research organizations, data, and computing centers

- Software services (authentication, search)
- Expert services (training, user documentation, help desk)
- European Open Science Cloud (EOSC): virtual environment for hosting, processing, sharing, and accessing research data across borders and scientific disciplines
 - Software services
 - Marketplace (federated catalog)
 - Provider Hub (documentation)
 - Expert services
 - Recorded lessons on marketplace resources
 - Help desk is for uses or suggestions
- Open Data Infrastructure for Social Science and Economic Innovations (ODISSEI): national research infrastructure for social science in the Netherlands
 - Software services
 - Data facility
 - Observatory
 - Laboratory
 - Expert services
 - Hub (social data science team)
 - Monthly online drop-ins for consultations
 - Data management expert guide
 - ASReview AI Software
 - Active learning in systematic review

Domestic U.S. providers

- NORC Data Enclave: secure research platform
 - Software services
 - Standard data services (security, linkage)
 - Research services (data deidentification)
 - Cloud products
 - Expert services
 - Metadata documentation
 - Data dissemination
 - Analysis and modeling
- ICPSR NACJD: official archive for data produced by the Bureau of Justice Statistics (BJS)
 - Software services
 - Discover: search by sponsor, product, format, restriction, type, topic

- Analyze data online (Survey Documentation and Analysis – SDA): view frequencies, crosstabs, correlation matrixes, subset, compute new variables
- Variable-level search (Social Science Variable Database – SSVD)
- Expert services
 - Contact email address for data deposit support (NIJ, OJJDP, BJS, Other)
 - Bibliography (search by authors, journals, sponsor), view all citations
 - Learning and data guides, resource guides, FAQs
 - Restricted data application forms, sample forms, requirements
- Coleridge Initiative Administrative Data Research Facility (ADRF): secure FedRAMP authorized research platform for access to and discovery of sensitive microdata
 - Software services
 - Operational reporting and resource provisioning
 - Data-sharing and provisioning
 - Secure data hosting and linkage
 - Expert services
 - Disclosure risk review for data exports
 - Webinars
- FSRDC: partnerships between federal statistical agencies and research institutions to provide secure data access environments
 - Provide opportunities for researchers to analyze use statistical data
 - Software services
 - SAP Portal for restricted data access from 16 federal statistical agencies
 - Secure Remote Research Environments for restricted data access for qualified researchers and projects
 - Microdata search (demographic, economic) and planning database
 - Expert services
 - Videos, consultations for restricted data use access
 - List of primary points of contact for each location
 - SHADAC/CDC
 - Access federal data to address state policy maker and federal employee needs
 - ResDAC/CMS
 - Compile and make data available
 - Support data applications
 - Conduct trainings
 - FSCPE/States
 - Assist other state agencies with population projections using public data

Appendix

Outreach Materials and Interview Guides

Federal Statistical Agencies

Initial Email Invitation from NORC

Subject: Request to Participate in an Interview about Models for a Data Concierge Service

Dear [NAME],

I am writing to ask for your assistance on a project sponsored by the National Center for Science and Engineering Statistics (NCSES). NCSES has contracted with NORC at the University of Chicago to develop models and tools for a data concierge service for a potential, future National Secure Data Service (NSDS). The concierge service models will be designed to assist researchers seeking access to data for evidence-building and research. This project supports the provisions of the 2018 Evidence Act and the recommendations of the Advisory Committee on Data for Evidence Building.

To fully understand the current support services available, we are contacting each of the Federal Statistical Agencies to setup individual discussions to learn more about the support they provide to data users for evidence-building; what they see as limitations in providing this support; and what enhancements they feel would best improve data access and sharing for evidence-based work. The data collected from this effort will help more clearly identify the “as is” data concierge support environment and development of additional data concierge models that account for your agency’s vision for future services.

Please designate a point of contact from your agency who you feel is best suited to meet with us to respond to the questions outlined below. Once we receive a reply with your point of contact (including contact information), we will work with that person to schedule a meeting.

Our goal is to complete this information gathering phase by November 10, so we’d appreciate a response as soon as possible. If you have any questions about this project, please contact Heather Madray, hmadray@nsf.gov. Thank you for your support in furthering the development of an evidence-based data sharing infrastructure.

Interview Protocol

- I. **Introduction and Purpose of the Data Concierge Project**
- II. **Questions to Guide the Discussion**
 1. How do people tend to find information about and access your data?
 2. Does your agency provide specialized assistance to potential users of your data in any of

- the following areas (if yes, please describe that service)?
- a. Assistance in refining the user's data requirements statement?
 - b. Data discovery?
 - c. Negotiating data access/acquisition agreements?
 - d. Data linkage methodologies?
 - e. Data quality assessments?
 - f. Any other areas that are not listed above?
3. Does your agency have a publicly available inventory of data potentially available for evidence-based projects? If yes, please provide a link in the chat.
 4. What are the legal authorities for your agency that govern access to your data?
 5. Is there a category of variables (e.g., person or business identifiers) that your agency generally considers too sensitive to share for evidence-based projects? If so, please describe.
 6. What are the typical reasons that potential users of your data might be denied access to those data?
 7. What suggestions do you have to mitigate the restrictions on sharing your data?
 8. For any data that you have provisioned to the Standard Application Portal (SAP), do you get user requests for assistance or requests for clarification on any aspect of those data? If so which areas?
 - a. What specific user services and/or tools do you provide? And why?
 - i. What are the barriers to providing additional services?
 - ii. What specific questions do you receive from users?
 - iii. Are you typically able to answer user questions?
 - b. Are there any other repositories where your agency data are accessed?
 - c. If you have not made data available to the SAP, is there a reason why?
 9. Are you staffed to a level to respond to all requests for assistance regarding your data? If not, what specialties/expertise would help?
 10. What specialized data services would you like to see as part of a future concierge infrastructure?
 11. What tools are needed to support a future concierge service?

Data Users

Organizational Contact Email from NORC

Subject: Request to Participate in an Interview to Inform the Development of Federal Data Concierge Service

Dear [ORGANIZATION NAME],

The National Center for Science and Engineering Statistics (NCSES) within the U.S. National Science Foundation has contracted with [Contractor] to develop models and tools for a data concierge service for a potential, future National Secure Data Service (NSDS). The concierge service models will be designed to assist researchers seeking access to federal data for evidence-building and research.

As part of this effort, [Contractor] is conducting key-informant interviews with a mix of data users. Your organization has been identified as a user of federal data, and we are asking your organization to

designate a point of contact who is best suited to meet with us to discuss their experience seeking access to federal data. Once we receive a reply with your point of contact (including contact information), we will send the individual an introductory email and will work with them to schedule a virtual meeting. If you have any questions about this project, please reply to this email or contact Heather Madray, hmadray@nsf.gov. Thank you for your support in furthering the development of an evidence-based data sharing infrastructure.

Initial Email Invitation from NCSES

Subject: Request to Participate in an Interview about Models for a Data Concierge Service

Dear [NAME],

The National Center for Science and Engineering Statistics (NCSES) within the U.S. National Science Foundation has contracted with [Contractor] to develop models and tools for a data concierge service for a potential, future National Secure Data Service (NSDS). The purpose of this study is to recommend approaches for providing a range of high-quality services to support data users who are pursuing evidence building research. As part of this effort, [Contractor] will be conducting key informant interviews with a mix of data users, including researchers from academic institutions, non-profit organizations, think tanks, state and local government, economic development organizations, minority serving institutions, and professional organizations. We will use these interviews to provide context about federal concierge services, identify challenges and limitations to provision of concierge services, and identify any additional information that may be useful in a future concierge service.

You were recommended to us by [name/organization] as someone who would provide an important perspective on this work. You will receive an email from [Contractor] 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 develop data concierge service models and improve support to data users pursuing evidence building research. The more viewpoints and input we have, the more we can ensure the concierge models meet the needs of a wide variety of data 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.

Follow-up Email from NORC

Subject: Follow-up on NCSES Request to Participate in an Interview about Models for a Data Concierge Service

Dear [NAME]

Recently, The National Center for Science and Engineering Statistics (NCSES) within the U.S. National Science Foundation sent you an email about a project exploring models for a future data concierge

service. NCSES has contracted with [Contractor] to develop models and tools for a data concierge service for a potential, future National Secure Data Service (NSDS).

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 [Contractor], about your work and the resources you use when accessing federal data. We are interested in your perspective regardless of whether you are familiar with an existing data concierge service. All interview responses will be kept confidential and reported in aggregate. [Contractor] will not share your interview responses with anyone outside our study team.

We are following up to coordinate a date and time for this interview that is convenient for you within the next two weeks. We expect the discussion to last 90 minutes. In the table below please indicate a few dates and times most convenient for you. We will then send you a calendar invite, including Zoom meeting link, for one of your preferred dates. When you respond with your date and time preferences, please also note your current time zone and your approval to record the Zoom meeting.

Date	Time	Time zone

If there are other data users whom you would recommend that we contact for this project, please provide their name and email address.

We are happy to answer any questions or address concerns you may have about the content or purpose of the interview. Please reach out to Seth Brohinsky with any questions (brohinsky-seth@norc.org) or Heather Madray (hmadray@nsf.gov).

We look forward to hearing from you!

Interview Participant Consent Information

1. **Why are we doing this Study?** We are conducting this research study to develop models and tools for a data concierge service for a potential, future National Secure Data Service (NSDS). A data concierge service would offer technical assistance to individuals seeking access to federal data.
2. **Who is funding this Study?** This study is sponsored by the National Center for Science and Engineering Statistics (NCSES) within the U.S. National Science Foundation.

3. **What would I be asked to do if I am in this study?** You will be asked to participate in a 90-minute virtual discussion with an interviewer from NORC about your work and the federal resources you use.
4. **Is participation voluntary?** Your decision to participate in this research is voluntary. You can stop at any time. You do not have to answer any questions you do not want to answer.
5. **Are there any risks in participating in this study?** There are no risks in participating in this research beyond those experienced in everyday life.
6. **What are the benefits of participation?** Your perspectives and insights are critical in developing data concierge service models and tools that meet the needs of a wide variety of data users and improve support to those pursuing evidence-building research.
7. **How long will it take for me to participate in this study?** It will take about 90 minutes to complete the interview.
8. **Who will have access to the results of this study and/or my information?** 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 or NCSES. The recording and transcriptions will be used for internal purposes only and will be a resource for the development of written recommendations or research reports. Everything you share will be kept confidential to the extent allowed by law.
9. **Who can I contact with questions?** Please contact Seth Brohinsky at 978-559-1472 with questions, complaints, or concerns about this research. If you have any questions or concerns about your rights as a research participant, please contact the NORC Institutional Review Board (IRB) Manager by toll-free phone number at (866) 309-0542.

Interview Protocol

Hello. My name is [NAME], and I am a [TITLE] in the [DEPARTMENT NAME] department here at NORC. We are conducting this work on behalf of the National Center for Science and Engineering Statistics within the U.S. National Science Foundation.

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 hearing about the kinds of experiences you have had when seeking access to federal data. We plan to interview about 15-20 individuals. Our goal is to gain a better understanding of the resources people use and experiences they have when seeking access to federal data as well as suggestions for developing a new data concierge service that would provide technical assistance. We anticipate the discussion will last around 90 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 [Contractor] study team or National Center for Science and Engineering Statistics. When we submit our final report of findings, your responses will be combined with the responses of the other participants.

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 recordings will be destroyed once the project is complete.

The consent form was sent prior to this interview. Do you have any questions about the consent form? Do you consent to be part of this study? Please acknowledge your consent in the chat.


Do you agree to have this conversation recorded? (Yes/No)

[IF YES:] Ok, the recording has started. Do you have any questions or concerns before we get started?

1. How would you describe your current role?
 - i. In your current role, what specific types of federal data do you typically access?
 - ii. In your current role, what types of non-federal data do you typically access?
2. If you use federal data, where do you typically search for federal data?
 - i. What search strategies do you use? Do your strategies depend on the data source?
 - ii. What websites do you use in your search?
 - iii. What types of interactions do you have with source data contacts when trying to access data? Do you tend to rely on interpersonal communication with source data contacts you are familiar with and/or have worked with before when searching for data?
 - iv. Are you aware of the federal Standard Application Process (SAP) (<https://www.researchdatagov.org/>)? Have you used it? If so, what are your thoughts? If you are not familiar with the SAP we will post the URL in the chat at the end of the interview, if you are interested in learning more about it.
3. In your role as a [policymaker, agency executive, practitioner, subject matter expert, [FILL FROM RESPONSE TO #1]] have you had a recent instance when you were seeking access to federal data with which you were unfamiliar? If yes:
 - i. What information did you need to better define the characteristics and source of the data?
 - ii. What information about the data (author, keywords, identifiers, etc.) helped you search?
 - iii. What questions did you have about the data as they pertain to your research question?
 - i. PROBE: Did you have a known dataset or variable in mind or were you exploring available data?
 - iv. Did you have resources available to you, such as a subject matter expert or online information, to draw upon to assist you? If so, please describe?
 - i. PROBE: How did these resources help you determine if the data would meet your needs?
 - v. How did each resource (Subject matter expert or online information) assist you in determining the fitness for use of the data?

- vi. Did any resources help you envision the types of analysis the data could support? How did you originally find out about these resources?
 - i. PROBE: Did any of the resources you used point you directly to the data?
- 4. Were the resources you used helpful or unhelpful? Can you say more about that?
 - i. Would you consider your search successful? Please discuss your experience.
 - i. PROBE: Were you able to locate relevant data?
 - ii. PROBE: Can you tell me more about what led you to not being able to locate the relevant data? Were there any barriers you faced?
 - iii. PROBE: If you were able to locate relevant data, were you able to access the data in a format that you could use?
 - ii. Which resources for discovery and access of data were the most helpful? Can you say more about that?
 - iii. Which resources were the least helpful in your efforts to discover and access data? Can you say more about that?
- 5. Next, I'd like to ask you about different kinds of help that may be available for helping you locate and access data. A data concierge service provides technical assistance in the discovery, access, and use of federal data. Have you ever used a data concierge service? If so, is there a specific data concierge service available to you? [IF NO, SKIP TO QUESTION 6]
 - i. In your current role, have you ever used a data concierge-like service? If so, please describe.
 - ii. What data or assistance were you seeking?
 - i. PROBE: Did you have a specific dataset or multiple datasets you were trying to access; were you trying to see what datasets are available for a given topic or research domain; or both?
 - iii. What specific questions did you have?
 - iv. What did you learn through your interactions with the service?
 - i. PROBE: Tell us about your experiences with the services provided. Did you receive any guidance or assistance from the data provider to gain access to the data?
 - v. How specifically did you access data through the service?
 - i. PROBE: If data were restricted or private, did you access them through a secure portal, download or other means?
 - vi. What were your general reactions?
- 6. Did you use other support resources other than those we have already discussed? (If yes) Can you tell me more about that?
 - i. What was that experience like?
 - ii. What aspects of the service worked well?
 - i. PROBE: Access, usability, anything else.
 - iii. What aspects of the service didn't work well?
 - i. PROBE: Were there questions the provider could NOT answer? If yes, what were those questions?

- ii. PROBE: Access, usability, anything else.
 - iv. What, if any, were the limitations of the service?
 - i. PROBE: Access, usability, anything else.
 - v. Thinking about future data needs, would you be likely to use this service again?
 - i. PROBE: Tell me more about that.
- 7. As a [ROLE] I'm interested in what specific concierge services you would find particularly useful. A recent government report from the Advisory Committee for Evidence-Building report imagined a concierge service could potentially: (1) provide detailed steps to assist data users in finding confidential data; (2) help develop a proposal to request access to confidential data; and (3) provide direction to specialized resources or services to navigate the data request.
- 8. As we imagine developing data concierge models that can help users identify appropriate data for evidence-based research, we may propose solutions that vary in their technical complexity and staffing requirements. We believe that there could be at least two models defined to offer these services. The first model, Model A, compiles the "as-is" support provided by federal agencies into one centralized web location. The second model is full-service model, Model B, that builds on the foundation established by the connector model to add features and tools to assist data users to discover, access and use federal data. We'd like your reactions and input on both.
 - a. Model A would provide users with general information about data assets with a focus on relying as much as possible upon existing agencies' contacts and processes. With this model you would have access to a list of agency contacts and the ability to email or call a help desk for assistance in discovering, accessing, and using federal data.
 - i. What are your first impressions/general reactions?
 - ii. What specific services, if any, would be useful?
 - iii. What, if anything, might not be useful about this kind of service?
 - b. Model B offers expanded services, more types of service providers, more customized tools, and increased leveraging of artificial intelligence (AI). With this model you would have access to additional guidance in refining data user requirements (through subject matter expertise), support for data discovery (through services like chatbots), negotiating data agreements for restricted data, and data linkage.
 - i. What are your first impressions/general reactions?
 - ii. What specific components, if any, would be useful in this model?
 - iii. What, if anything, might not be useful about this kind of service?
- 9. Are there any additional enhancements, changes, or missing features of a data concierge service you recommend we address?
- 10. Are there other resources we have not yet discussed that you find particularly useful when looking for federal data? If yes, please describe the resources and how you use them in your work.
- 11. Are you aware of other federal data users who can provide unique insights into their experience accessing federal data that you recommend we include in this study? If so, can you provide their contact information for us to reach out?



12. 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?

On behalf of the National Center for Science and Engineering Statistics, I want to thank you for your time today. I appreciate the important insights you have shared. If you have any questions or think of anything else, please reach out.