Interagency Council on Statistical Policy
Leaders of the United States Federal Statistical System

Mathematica
Progress Together

Data Usage Platform (DUP) Prototype

Submitted August 29, 2025

Submitted to:

National Center for Science and Engineering Statistics
2415 Eisenhower Avenue, Suite W14200
Alexandria, VA 22314
Phone: (703) 292-8780
Fax: (703) 292-9092

Submitted by:

Mathematica
1100 First Street, NE, 12th Floor
Washington, DC 20002-4221
Phone: (202) 484-9220
Fax: (202) 863-1763

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, NSF, or their government partners. Please send questions to ncsesweb@nsf.gov. This product has been reviewed for unauthorized disclosure of confidential information under NCSES-DRN25-058.

Contents

Overview of DUP Prototype

Purpose

As part of the National Secure Data Service Demonstration (NSDS-D) project, the National Center for Science and Engineering Statistics (NCSES) contracted with Mathematica and datHere to design and build a prototype Data Usage Platform (DUP). The DUP is a shared-service framework that enables federal agencies and the broader data community to better understand how federal data assets are used in research, news media, policymaking, and publicly available reports.

The project spanned three phases: Discovery and User Research, Design and Development, and Communication and Documentation. Together, these phases showed how a user-centered, artificial intelligence (AI)-enabled platform can curate usage data and present those data through an accessible dashboard. The prototype addresses several information gaps from the scope of work, including which features users need, how usage data can inform tiered access and transparency, and what technologies can reliably generate aggregated usage metrics.

Vision

We envisioned the DUP as a government-wide shared service and potential cornerstone of the NSDS. Built to be sustainable, standardized, open source, and interoperable, the DUP prototype demonstrates a flexible and nimble approach that can be used to summarize usage data across federal agencies and different contexts. Although existing efforts such as the Democratizing Data (DD) initiative pilot served as an informative reference, our DUP prototype project deliberately explored alternative technologies, use cases, and other innovative approaches to think outside the box and surpass traditional data asset frameworks. In that sense, this effort followed a start-from-scratch approach that prioritized interoperability over agency-specific tailoring, adhering to federal requirements with a modern, user-centered, state-of-the-art backend.

Background

We built on lessons from the DD pilot dashboards, which were launched in 2016 to support the Commission on Evidence-Based Policymaking. The DD initiative demonstrated the feasibility of tracking federal data usage in academic literature and visualizing it through dashboards. However, it also revealed gaps in data ingestion, metadata consistency, and user engagement. The goal of the DUP prototype project was to design a state-of-the-art solution that addresses gaps from the DD Initiative based on user feedback and engagement that we collected throughout the research and development phases.

Stakeholder engagement overview

We engaged two stakeholder groups throughout the project: **Data users** (federal and state staff, researchers, journalists, and policymakers) and **data owners and creators** (analysts and leaders at federal nonstatistical agencies). Generally, data users search for specific federal data assets and have a good understanding of well-known federal data assets but are inquisitive about what other data assets are available and who are the authors of highly cited data assets. Data owners want to track how their

assets are used and by whom. These perspectives shaped our approach and requirements to constructing usage metrics.

Project phases overview

In **Phase 1: Discovery and User Research**, we interviewed and conducted usability tests with 51 participants from the two broad stakeholder groups across sectors to understand behaviors, needs, and constraints, producing personas, journey maps, and a gap analysis that informed requirements. In **Phase 2: Design and Development**, we implemented a modular ingestion pipeline and an iterative, responsive dashboard, pairing rule-based methods with large language models (LLMs) for tagging and quality assurance (QA). In **Phase 3: Communication and Documentation**, we created user-ready materials and technical documentation to support awareness, reuse, and future adoption across agencies.

Phase 1: Discovery and User Research

Goals and objectives

Phase 1 of the DUP project centered on foundational discovery to inform the design of a prototype. We focused on three core information gaps: (1) what data, features, and functions different users need; (2) how usage profiles can inform tiered access, transparency, and open data requirements; and (3) what users find useful in existing usage dashboards like the DD pilot dashboards.

To capture a wide range of perspectives beyond those of federal data owners, we included nonstatistical federal agencies alongside researchers, journalists, advocacy groups, academics, and policy staff. This wide-ranging outreach allowed us to explore how public data assets are interpreted and applied across different domains, helping us identify opportunities where the DUP could add value to broader audience types.

This phase generated a detailed set of user-informed feature requirements and data priorities that served as direct inputs into Phase 2 design activities.

Methodology and outreach

Methodology

We employed three types of user research methodologies: user interviews, focus groups, and usability testing. From these approaches we intended to gather a wide range of qualitative insights from diverse stakeholders. User interviews served as the primary method based on scheduling flexibility and in-depth focus, allowing for tailored conversations that revealed nuanced needs and perspectives. Focus groups were used sparingly, as they proved logistically challenging to organize across main user groups and roles. Usability testing in this phase was limited to evaluating the existing DD pilot dashboards, helping identify design strengths and areas for improvement.

While we waited for fast-track approval from the Office of Management and Budget (OMB), we began our user research with federal staff (exempt from OMB) to reduce delays, gather early insights, and guide our eventual test participant recruitment. Once clearance was obtained, we expanded our research activities to include a diverse set of participants: federal chief data officers, information

technology (IT) staff, state and local agency representatives, academic researchers, journalists, and open data advocates. Each session followed an OMB-approved semi-structured protocol tailored to the participant's role and focused on key themes such as data publishing practices, usage tracking, challenges in accessing federal data, and visions for public data transparency.

Outreach and research participants

We identified potential participants through NCSES and partner networks, referrals, and online outreach, with a particular emphasis on nonstatistical agencies and academic researchers. We also included several federal IT staff to understand challenges with existing technology solutions and identify future recommendations for the prototype development.

Overall, we contacted 150 stakeholders for Phase 1 and completed sessions with 51 participants:

- 35 participated in interviews
- 14 participated in focus groups
- 2 participated in usability tests

Some of these participants also participated in Phase 2 usability tests. See Exhibit 1 for a breakdown of all 51 participants.

Exhibit 1. User research participant groups

User group	Number of participants	Description	Example data sets or sources created/used
Data owners and creators			
Federal agency chief data officers, leadership, and analysts	23	Employees of federal agencies who are involved in data collection or construction efforts that result in federal data sets, some of which are shared publicly	U.S. Census American Community Survey, BLS labor data, NCSES surveys, NSF awards, national accounts from Bureau of Economic Analysis, NCES Integrated Postsecondary Education Data System, WARN Act, ACF restricted data, Collective Bargaining Agreement data
Federal IT or developer staff	6	Technology leads who implement federal data dashboards, technology applications, and other tools	Note: NCSES and BLS developers implemented data usage tracking for their own agency-specific data sets.
Data owners and creators / data users			
State and local agency staff	6	State and metropolitan government agency staff who collect federal and local data to evaluate their programs	EPA, U.S. Census American Community Survey, BLS labor data, HUD housing data, DOL workforce data, other local survey data
State-run federal program managers	6	State officials involved in implementing federal programs (SNAP, TANF, Medicaid, and Medicare) who collect local data and use those data to evaluate programs	TANF, SNAP, LIHEAP, Medicare and Medicaid, U.S. Census American Community Survey, HEDIS, NDACAN data, other local survey data

User group	Number of participants	Description	Example data sets or sources created/used
Data users			
Journalists	3	Journalists or other media academic professionals who use federal data in their reporting	U.S. Census American Community Survey, BLS labor data, FBI and local police crime statistics, CDC and CMS data sets, IRS Form 990, federal court case statistics, federal and local election contributions data, CMS hospital data
Researchers and open data advocates	7	Researchers who use federal data in their work and have a broad understanding of usage statistics	U.S. Census American Community Survey, BLS labor data, CDC National Vital Statistics Program, Medicaid and Medicare claims data, USDA agriculture data, HUD data

ACF = Administration for Children and Families; BLS = Bureau of Labor Statistics; CDC = Centers for Disease Control and Prevention; CMS = Centers for Medicare & Medicaid Services; DOL = U.S. Department of Labor; EPA = Environmental Protection Agency; FBI = Federal Bureau of Investigation; HEDIS = Healthcare Effectiveness Data and Information Set; HUD = U.S. Department of Housing and Urban Development; IRS= Internal Revenue Service; IT = information technology; LIHEAP = Low-Income Home Energy Assistance Program; NCSES = National Center for Science and Engineering Statistics; NCES = National Center for Education Statistics; NDACAN = National Data Archive on Child Abuse and Neglect; NSF = U.S. National Science Foundation; SNAP = Supplemental Nutrition Assistance Program; TANF = Temporary Assistance for Needy Families; USDA = U.S. Department of Agriculture; WARN = Worker Adjustment and Retraining Notification.

Analysis

User research analysis

To synthesize insights from our user research, we conducted a thematic analysis of qualitative data collected through user interviews and focus groups. User interviews and focus groups aligned around the same core themes, and their outputs were merged into a single coded data set to ensure consistency and depth across participant types. Results from usability testing, which focused specifically on the existing DD dashboards, were analyzed separately. Findings from usability sessions were translated directly into design requirements and were not included in the broader thematic coding.

We used thematic analysis to effectively identify patterns across qualitative data. We coded user research notes based on predefined themes from the interview and focus group protocols such as the types of data participants use or create, data findability, data usage tracking, and dimensions of data use. In addition to these structured themes, the team remained open to emergent patterns, allowing new thematic codes to surface organically from the data. We compiled all coded notes into a Mural board and collaboratively analyzed them through an affinity diagramming exercise with key project stakeholders. Using this hands-on, visual approach, we identified recurring patterns, built consensus around findings, and developed a shared understanding of user needs and priorities.

User personas

From the analysis of user research outputs, three primary personas emerged:

- **Federal nonstatistical agency staff.** These users track both internal and external data usage to support funding justification and enhance transparency. Their primary needs are to understand how their data are accessed and used across different sectors.

- **State agency staff.** This group is primarily concerned with integrating federal data sets into localized programs such as Temporary Assistance for Needy Families, the Supplemental Nutrition Assistance Program, or Medicaid. Their use cases emphasize the need for interoperability and contextual relevance of data.
- **Data journalists.** These users require intuitive, topic-based search capabilities and direct access to data assets to support storytelling and investigative reporting. Their priorities include ease of discovery and clarity in data presentation.

Use cases

User research revealed several promising use cases for the DUP that extend beyond internal federal agency needs. Participants across federal, state, and local agencies, as well as journalists and researchers, highlighted the platform's potential to increase data awareness by centralizing access to both common and lesser-known data assets, helping users navigate the fragmented landscape of federal data sources.

The DUP was also seen as a valuable tool for tracking long-term data usage trends, enabling users to identify popular data sets and emerging topics through citations and media mentions. Agencies expressed interest in comparing data usage with similar nonstatistical federal agencies, using the platform to benchmark their performance and visibility. Additionally, the ability to track access and usage, including downloads and user interactions, was viewed as essential for understanding the full life cycle and impact of public data assets.

Participants emphasized the importance of measuring contextual impact, noting that usage statistics alone cannot support funding requests or inform decisions to scale back underused programs. The DUP could also help agencies reduce the burden of data collection by aligning efforts with actual user demand. Finally, by offering insights into who is using the data and how, the platform can foster greater trust in federal data and support transparency across sectors.

Key findings

Phase 1 user research provided key insights into how nonstatistical federal, state, and local agencies; data journalists; and researchers interact with public data and what they need from a future state DUP. These findings directly address information gaps identified by NCSES and highlight opportunities to expand the platform's use across all audience groups.

- **Supporting evidence-building across sectors.** Participants emphasized the need for centralized access to federal data tracking, noting that fragmented agency-specific platforms usually hinder data discoverability. Journalists and researchers expressed interest in topic-based search and metadata-rich interfaces, while federal and state agencies sought tools to compare usage across peers. These findings suggest that the DUP should support diverse use cases across sectors, including media reporting and academic research.

- **Leveraging usage data for transparency.** Agencies currently track basic access metrics such as downloads and page views but lack insight into how data are used post-access. Participants highlighted the value of usage data in justifying funding, improving transparency, and informing decisions about data accessibility. Usage data could help agencies prioritize data collection efforts and identify underutilized data sets needing better outreach or documentation.
- **Going beyond usage statistics.** Although participants agreed that usage statistics are useful proxies for data impact, several agency staff and policymakers cautioned against relying solely on them. They described instances of data sets with low citation counts that may still have significant policy relevance. Usage data can help identify gaps, support resource allocation, and guide program development but should be complemented with qualitative insights.
- **Building public trust.** Participants viewed data usage tracking as a way to enhance public trust in federal statistics. By making data usage visible and understandable, the DUP could support transparency and engagement, aligning with the goals of a future NSDS.
- **Technologies and best practices for usage tracking.** The research identified several technologies that were used to track usage across agencies, including Tableau, Qlik, Google Analytics, OpenAlex, and Google News SerpApi. We identified these open-source and low-cost tools as possible technologies for a future state of the DUP because of their scalability and replicability. Manual processes such as keyword updates and CSV downloads remained common processes for tracking internal federal data usage, highlighting the need for automation and sustainable infrastructure.
- **Streamlining data asset metadata.** Participants highlighted how inconsistent naming of metadata across agency platforms creates challenges to search for and track data assets. The DUP offers an opportunity and an incentive to streamline naming conventions and standards for data asset metadata across agencies.
- **Providing responsive intuitive dashboards.** Participants noted the DD dashboards were difficult to navigate and lacked responsive design, meaning that they did not automatically resize and adjust the layout based on the size and orientation of the user's screen. The DUP must not only address substantive user requirements but also provide responsive and user-centered ways to access information on data asset usage.

Lessons learned

The Phase 1 research process offered valuable insights into user needs and the operational challenges of implementing the DUP.

Workflow process and governance

The OMB fast-track process, though intended to streamline approvals, still required multiple rounds of internal review that delayed engagement with non-federal participants. This highlighted the importance of early planning and flexible coordination to ensure that user research activities could be completed on time and inform the design phase.

If required, future user research and engagement activities should start before OMB clearance, when feasible, to collect rapid-cycle feedback and implement platform updates. When clearance is required, breaking submissions into smaller, method-specific packages (interviews, focus groups, usability testing) could reduce bottlenecks and contribute to a leaner user engagement process.

Stakeholder engagement

Our outreach initially focused on federal data leaders, then broadened to smaller federal and state agencies that proved more responsive and had pressing needs for usage tracking. Focus groups were harder to schedule and less productive than expected, while one-on-one interviews yielded richer insights.

Outreach for future DUP user needs sensing should prioritize interviews as they are a reliable and effective way to capture detailed, actionable insights.

Federal data usage and web metrics

While web metrics such as Google Analytics are widely used across agencies to monitor page views and downloads, we chose not to integrate these metrics into the prototype, as doing so would have too complex for this endeavor.

Given the scope and technical constraints of Phase 1, the DUP prototype concentrated on federal-level data, intentionally excluding state and local data usage tracking. This decision was not due to a lack of interest or relevance for non-federal data, as user research revealed strong potential use cases among state and local organizations, many of which expressed a desire to better understand how their data are accessed and used.

Future iterations of the DUP can build on the existing foundation to incorporate broader data sources, including state and local data sets, and more advanced usage tracking capabilities that were not included in this prototype.

Implications for Phase 2 and beyond

The insights gathered in Phase 1 discovery informed concrete design, data, and technology requirements for the DUP prototype. They also pointed to broader needs that a DUP should address: a scalable, inclusive platform that supports diverse user groups, integrates with smaller agency systems, and builds public trust through transparent usage metrics. Future development should continue to prioritize user engagement, technological flexibility, automation, and accessibility to ensure the platform meets the evolving needs of its users.

Phase 2: Design and Development

Goals and objectives

Phase 2 translated Phase 1 requirements into a working prototype that curates, validates, and visualizes usage of federal data assets. Throughout this phase, we evaluated feasibility and trade-offs across multiple dimensions: automated versus manual curation of data assets, front-end frameworks, metadata standards, and AI-enabled tools. We addressed four information gaps: (1) which

open data tools and methods reliably produce aggregated usage data; (2) what resources support a flexible, sustainable platform; (3) how usage statistics can indicate the health of the federal statistical system; and (4) how a shared service can build public trust.

Our approach centered around AI-forward, open-source solutions that scale, reduce agency burden, and align with the NSDS vision. We organized the work into two coordinated streams:

- **Workstream 1:** We built a modular data pipeline to ingest and process usage data that aimed to minimize the need for agency staff intervention while ensuring quality and flexibility. The pipeline identifies assets, extracts publications, tags asset references, runs automated AI-assisted QA, and loads results into a database that the dashboard could query in real time. We partnered closely with the front-end development team, so the pipeline output matched user interface needs.
- **Workstream 2:** We focused on designing and developing the prototype's front-end platform, driven by iterative testing with end users. Design priorities included intuitive navigation, topic-based search, metadata clarity, and responsive visualizations of usage metrics.

Workstream 1: Data ingestion pipeline

In Workstream 1, we curated a set of federal data assets and built a modular pipeline that (1) extracts candidate publications across channels, (2) tags asset references, (3) runs automated QA, and (4) loads results into a relational database that the dashboard can query. The modules are independent so new sources or models can be added without rearchitecting the system.

Curating federal data assets

Before developing the data ingestion pipeline, we curated a set of 60 high-priority federal data assets from 13 agencies to demonstrate breadth while keeping scope feasible. We prioritized assets that were likely to appear across multiple publication types, spanning agencies, and including both public and restricted-use data (Standard Application Process, SAP). We identified our data assets following three main steps:

1. We identified potential assets by scraping asset metadata from Data.gov using the CKAN API.
2. We implemented a streamlining process that grouped data assets with similar names published by the same agency; excluded data assets formatted only as images, PowerPoints, or HTML files that are unlikely to be used in evidence-building capacities; and generalized variants under unified asset names. This left a set of more than 2,000 assets and information on their common abbreviations and sponsoring agencies.
3. From this set, we collaborated with NCSES to identify the 13 priority sponsoring agencies and the 60 priority assets sponsored by those agencies.

Standardizing asset names improved metadata quality and made downstream identification more reliable. Searching for publications with mentions of potential data assets works best when each asset has a single, canonical name that is general yet sufficiently distinctive. Too much specificity increases false negatives, and too little context increases false positives. For example, the American Community

Survey includes many modules; querying each one separately can miss relevant publications due to over-specificity. By contrast, "Consumer Price Index (CPI)" is a common term worldwide; flagging "CPI" without context on the agency or country can surface unrelated mentions. Standardization substantially increased successful identifications, though further room for refinement remains.

Extract-transform-load (ETL) pipeline

The ingestion pipeline was organized into eight discrete modules that ran in three stages and supported an automated ETL process. The modular design of the ETL pipeline supports scaling and modifying individual components independently. This approach enables future integration of additional sources as well as new tagging approaches.

Stage 1: Extraction

In the first stage, extraction, we implemented three modules to scrape academic, legislation, and news publications that might contain mentions of the 60 priority data assets identified in our data curation process. Each module uses a different set of scraping tools, download processes, and validation steps to identify and download publications that reflect differences in how data about each publication type are tracked and made available.

The academic publication scraping module relies on the OpenAlex API to access a detailed database of metadata about publications. Notably, OpenAlex is a free and open catalog of scholarly research, offering a fully-fledged API with generous request limits. The OpenAlex API supports searching for publications based on details including publication date, publication country, open-access status, and keywords. The API also provides lists of locations where the full text of articles can be found. This allows the pipeline to systematically search for full text of open-access publications that can be downloaded directly. Given the richness of the OpenAlex data, we were able to apply several filters to the API queries to improve the likelihood of identifying relevant results. For example, we limited results to articles to avoid retrieving entire books and limited publication language to English.[1] Once the text is downloaded, the module validates the quality of the text for issues and to ensure the text is in English. If the text is corrupted, the module attempts to retrieve it from a different source.

The news publication scraping module employs several APIs (including News API and DuckDuckGo) to identify publications from common news sources and systematically search for their full text. Approximately 30 percent of news publications we collected did not contain full text due to some news sources offering subscription-only access. To address this issue, we explored options to identify similarities between article headlines and summaries to group articles that are likely to cover similar stories and, therefore, may reference the same assets.

The legislation scraping module is simpler and more streamlined because it relies on the Legiscan API, which makes identifying and accessing relevant legislative text straightforward.

[1] This is necessary because OpenAlex uses abstract text to determine publication language, and many non-English scientific publications print abstracts in both English and their primary language, so articles identified as English may not actually be.

Stage 2: Transformation

In the second stage, transformation, three additional modules clean the extracted data and use rule-based and LLM-based approaches to identify and validate mentions of data assets:

1. The text extraction and cleaning module streams text from validated documents, cleans it following standard best practices, and outputs a machine-readable text file.
2. The tagger module reads in the machine-readable text and uses a three-step tagging approach to identify mentions of federal data assets. This approach comprises rule-based matching for exact name matches, fuzzy matching to catch variants and misspellings, and an LLM-based[2] tagging model to infer context to tag mentions that would have gone otherwise untagged. The tagger module invokes the LLM only when rule-based confidence was low to right-size the use of more expensive and computationally demanding resources.
3. The automated quality analysis module employs an LLM-as-a-judge model to validate less certain results from the tagger.

Loading

We load cleaned metadata and tags into a MariaDB relational database with clear entities (asset, publication, mention/use). Read-only access protects data integrity while enabling real-time views. To improve performance of dashboard elements, we also added and tuned caching mechanisms.

Annotation and QA

To support reproducibility and minimize reliance on manual annotation to identify and validate mentions of data assets, the team implemented a staggered and flexible annotation approach. We selected initial documents through targeted querying of the OpenAlex API to maximize the likelihood of relevant data asset mentions. Human annotators then applied detailed guidelines to (1) identify mentions of federal data assets in the text; (2) codify which specific asset was being referenced, even when the same source appeared under different names; and (3) confirm whether the asset was used directly in the analysis or cited only tangentially for a subset of documents. These annotations informed subsequent tagging rules and Python scripts that scaled the process, and they provided a benchmark for evaluating the performance of the pipeline's LLM components.

Concurrently, we developed an automated LLM-as-a-judge QA workflow, where we had a different foundation LLM (ChatGPT) provide QA on the tags generated through our hybrid rule-based and LLM (Claude) tagger. In a sample evaluation, this automated QA produced results that closely matched human judgments, suggesting a viable path for scalable QA. Notably, this approach does not fully replace human input, and further testing is required when scaling up the approach to a broader set of data assets.

[2] We explored different types of LLMs, including open-source ones such as FLAN-T5, BART, and Llama, and concluded that the most performant and scalable models are currently proprietary foundation models (such as Claude and ChatGPT). We also experimented with different prompting strategies to fine-tune model performance.

AI integration

AI plays an integral role in identifying data assets, particularly for those where context is key for identification. However, though powerful tools, LLMs have limitations. They vary in quality, accuracy, use cases, cost, and efficiency. Because of this, we assigned human reviewers to validate the results of all stages of the tagger module, including the rule-based tagging, zero/few-shot LLM tagging, and automated QA (LLM-as-a-judge). We discuss some of these AI-driven approaches in more detail below.

LLM tagger. Initial evaluation efforts of the ETL pipeline's zero/few-shot tagging showed mixed results using Anthropic's Claude Haiku 3 and Haiku 3.5 models.[3] Importantly, the LLM tagger deals with a more complex task, as it only receives (a) text that has weak fuzzy matches or (b) text with no rule-based matches but with specific keywords that signal the use of data. This module has ample room for improvement by using more powerful models as they become available at lower price points per 1,000 input tokens. For the prototype, the LLM tagger was therefore used selectively—applied only in low-confidence cases where rule-based methods were unlikely to succeed—and its outputs were subject to additional review. This approach allowed the team to take advantage of the LLM's strengths while minimizing the risks of false positives and unnecessary cost.

Automated QA. The automated QA module uses an LLM-as-a-judge approach, guided by a carefully designed prompt that provides the publication text, identified mentions, and instructions to verify them; assess "direct use;" and return structured JavaScript Object Notation (JSON) outputs. After multiple prompt iterations and validation checks, the system was tested on 500 publications, showing strong performance: results were consistently well-formatted, aligned with human judgment 97 percent of the time, and effectively flagged false positives from weaker tagging methods. Importantly, the module also distinguishes between mere mentions of federal data sets and their direct use in analysis or argumentation, something rule-based methods cannot do, further demonstrating the potential of LLMs to automate nuanced tasks that previously required human review.

Many publications mention federal data assets tangentially. For example, an author may mention multiple data assets even when only one is used directly, or they may name a data asset in a context that does not entail direct use. Differentiating between a mention and direct use might be appealing for users of the DUP, and automating the process is not possible using rule-based methods. Therefore, promising results from the LLM tagger and the LLM-as-a-judge approaches offer a path toward automating tasks that could previously not be feasibly applied at scale.

Workstream 2: Design, development, and testing

In Workstream 2, we designed, developed, and tested a responsive, user-centered dashboard prototype. Priorities included allowing for modular visualizations, the ability to implement user-logins and role-based access to features, easy content authoring and management for text-based pages, the

[3] Anthropic's Claude Haiku models are fast and lightweight, optimized for speed, affordability, and real-time responsiveness, features particularly useful for high-volume tasks like ours. However, they are estimated to be about an order of magnitude smaller than GPT-3 in parameter size, which makes them more efficient but with less raw reasoning capacity.

ability to explore extended features like the community forum and AI chatbot, and seamless integration with data generated through the back end.

Design process

The design process included two iterative rounds of early- and late-stage design phases.

Early-stage design. We translated Phase 1 requirements into early-stage wireframes, which were essential to the user-centered design process. These grayscale layouts allowed the team to focus on navigation, accessibility, and clear data visualization. Their simplicity made them easy to update based on user and stakeholder feedback, supporting rapid iteration on the design and usability testing. Wireframes helped validate core design decisions and ensured the platform's structure aligned with user needs from the start. We included these early-stage wireframe designs in the first round of usability testing, enabling the project team to make quick updates and evaluate key features for the next, higher fidelity design iteration.

Late-stage mock-ups. We converted early-stage wireframes into high-fidelity mock-ups to provide guidance for developers and support additional usability testing. These designs reflected NSDS branding and offered a realistic preview of the platform's interface and features. We updated these mock-ups iteratively based on usability testing results, stakeholder input, and technical reviews, which helped validate core elements such as responsive layout, topic-based search features, and impact metrics across academic, news, and government sources. The mock-ups supported a user-centered, iterative design approach and were used for the second round of usability testing.

Dashboard development

We identified specific technical requirements for a dashboarding tool: (1) enabling quick and easy development of the DUP dashboard and other user-facing features, (2) providing tight integration with the DUP's back-end processing (see Workstream 1 above); (3) allowing for easy iterative updates as usability testing provided insights for refining the prototype; and (4) providing a robust, secure, scalable, and extensible platform for a future production-grade system. We chose Drupal for its no-code/low code capabilities, foundational out-of-the-box functionality for building applications, robust application framework, and open-source community support.

Usability testing

Our usability testing findings suggest that usage statistics paired with contextual information and clear visualizations can offer meaningful insights into the health and progress of the federal statistical system. Participants saw value in tracking trends, identifying high-impact data sets, and understanding how data are used across sectors. These insights support the DUP's potential as a shared service that can enhance transparency, build public trust in federal data, and contribute to broader engagement with a future NSDS.

Participants and method. We conducted three rounds of moderated, 1:1 think-aloud usability tests, recorded sessions, and thematically coded feedback. Our designs progressed from wireframes to high-fidelity mock-ups to early prototypes with sample data, with features added or refined after each round. Participants were recruited from individuals we contacted during Phase 1, referrals, and targeted

searches. The final sample (N = 24) included nine federal data owners or creators, 12 academic researchers, two data journalists, and one policymaker, providing varied perspectives on platform usability.

We analyzed feedback from usability testing as it related to core requirements and components of the DUP.

Purpose of tracking federal data usage. Overall, testers responded positively to the concept of a centralized dashboard to track how federal data are used in academic research, media, and policy reports. However, many testers initially misunderstood the platform's purpose when starting their sessions, often assuming that the DUP was a data repository rather than a tool for analyzing data usage. Based on user feedback, the dashboard component of the DUP was branded as the Data Usage Explorer (DUE). This highlighted the importance of clear messaging on the home page and intuitive navigation to help users quickly understand the platform's purpose and value (see Appendix A1. Home Page).

Dashboard interface. The dashboard interface was generally well received, with testers appreciating its clean layout and inclusion of various filters and visualizations. The most valued features included the publication details table and the visualizations on usage of data assets over time which testers saw as essential for finding data assets of interest and understanding their impact. Testers appreciated the ability to explore data usage by domain: academic, legislative, media, and reports (see Appendix A2. Dashboard Data Visualizations).

Data asset verification feature. Testers responded positively to the "Confirmed Use" feature pop-up, which allowed them to verify whether a data asset was used properly in a publication. Although initial understanding varied, most testers appreciated the concept once they explored the feature, noting that it added a layer of credibility and QA to the dashboard (see Appendix A5. Community and Data Usage Feedback Integration). Agency staff saw its value in validating data references, though they acknowledged the effort required to maintain it. Modal pop-ups throughout the dashboard, such as those for publication details and data asset metadata, were also well received. Testers found them informative and helpful for exploring context without leaving the page, especially when the pop-ups included full citations, author information, and links to source data (see Appendix A3. Data Assets Referenced in Publications and A4. Expanded Metadata and Pop-Ups).

Data asset usage feedback. The data usage feedback feature, which allows testers to share how they have used specific data sets and provide their perspective on their impact and relevance, received mixed reactions. Although some saw its potential to capture qualitative insights and demonstrate impact, other testers were unsure of its purpose or expressed concerns about how such feedback would be managed. Despite this, the idea of surfacing real-world examples of data use was seen as a valuable addition to the platform (see Appendix A5. Community Forum and Data Usage Feedback Module).

Community forum. Designed to foster collaboration and discussion among users, the forum received varied responses. Some participants found it useful for sharing knowledge and connecting with others, while others questioned its necessity given the existence of more established platforms. Still, the idea of a space for data users and providers to interact was seen as aligned with the platform's broader goals of transparency and engagement. Testers viewed the potential for agencies to participate in discussions

and respond to user questions as a way to build trust and improve data stewardship (see Appendix A, A5. Community Forum and Data Usage Feedback Module).

Lessons learned

Usability testing outreach and engagement

Reengaging Phase 1 participants proved effective for validating earlier insights, while outreach to new groups such as researchers from doctoral universities with moderate research activity expanded perspectives for the prototype. Recruiting journalists was difficult, leaving a gap in evaluating news-related features. Academic and federal contacts were easier to engage but became harder to sustain across multiple rounds.

Future outreach could benefit from broader engagement strategies. Social media offers scalable ways to reach diverse user groups, including journalists and early-career researchers. Different types of incentives may also help increase participation, especially among users with limited time. These approaches would support more inclusive and efficient recruitment, ensuring the DUP continues to reflect real user needs.

Technology and methods

AI chatbot. We explored the potential of an AI-enabled chatbot to enhance user engagement and experience. An AI chatbot also has the potential to reduce burden on the NSDS data concierge by allowing users to ask questions about data assets that would otherwise require manual analysis. Research and testing in this area also revealed some useful technical insights:

- An AI chatbot performs best when combining multiple retrieval-augmented generation (RAG) techniques. SQL-based queries handle numeric or metadata-heavy requests (for example, “What were the top 5 data assets referenced in 2016?”), while semantic search with a vector database retrieves relevant results even when user terms do not match exactly (for example, “What resources relate to farming?”). Using both approaches ensures accurate and flexible responses.
- Agent-driven processing improves chatbot efficiency by assigning specialized agents to each RAG technique under a coordinating “management” agent. Testing showed that multi-agent setups outperformed single-agent ones, delivering more dynamic, context-aware responses and finer control over which data the chatbot emphasizes.

ETL pipeline. We tested varied approaches for obtaining publication text and for tagging mentions of federal data assets that balanced coverage, accuracy, and cost.

- Titles and summaries of publications are unlikely to be sufficient to accurately identify mentions of federal data assets that are used in publications. However, full text is often difficult to scrape, with difficulty varying by publication type. Legislation may be easiest, while open-access academic publications are slightly more challenging, and news article text is often carefully paywalled. As a result, coverage across domains is inherently uneven, so it is important that a

future state DUP is able to communicate limitations of the data clearly so users can contextualize data usage takeaways.

- Placing strong restrictions on search parameters for academic publications that may contain mentions of federal data assets substantially improves the probability of identifying mentions within full text. Despite these restrictions, the universe of academic literature the pipeline identifies for review remains very large. To speed up this process, a future state of the DUP should lean into the prototype's cloud-native approach and leverage parallelization, with multiple processes identifying and annotating relevant publications simultaneously. Similar approaches may be used for scraping state reports from state websites, which may not have APIs and, due to variable formatting, require a human-like web crawler to identify and access relevant information.
- A hybrid rule-based and LLM approach for tagging data asset references optimizes runtime and model accuracy. State-of-the-art proprietary LLM models are currently the most accurate for identifying and validating mentions of data assets. Although proprietary models require a paid license and fee per query, the multi-stage design of the tagger and explicit limits on per-query costs help keep costs manageable. Costs may also decrease as high-performing open-source LLMs become available in the future.
- In a sample evaluation, an automated LLM-as-a-judge approach, where ChatGPT rated the data asset tags generated by Claude in the ETL pipeline, achieved similar performance to human annotators, suggesting a viable path for scaling up QA at minimal human burden.

News data sources. Although news coverage is an important indicator of data asset use, capturing it systematically requires creative technical strategies and clear communication of methodological limits. Using existing APIs, we were able to retrieve full text for roughly 70 percent of identified news articles. To address gaps where text was unavailable, we began testing two approaches:

- Grouping strategies that cluster articles by title, summary, and publication date. This approach allows tags identified in one article with full text to be treated as "potential tags" for other articles in the same group. This method improves coverage, but it is important to note that it does not constitute imputation; rather, it offers a heuristic for inferring likely references without overstating accuracy. Potential tags are not currently stored in the system, but a future state DUP could consider how to display the difference between confirmed tags and potential tags so users understand the varying levels of certainty.
- Initial attempts to use LLMs to retrieve missing article text identified several additional challenges. Some sources, especially major outlets, restrict bot crawling or copyright-protected content, preventing reliable full-text retrieval. Although LLMs could summarize articles and extract potential data set mentions, this carries risks of fabricated output and unverifiable results. Alternative approaches, such as robust crawler indexes or custom-built scrapers, may provide better long-term solutions.

Workflow process and governance

The workflow alignment during DUP development highlighted collaboration strengths and challenges of scoping an ambitious work plan under a constrained schedule. Because design and development proceeded almost simultaneously to allow for iteration and initial proofs of concept, the teams had to make real-time decisions about which pieces of user feedback could be incorporated without jeopardizing delivery timelines. This created inevitable trade-offs in deciding which design refinements to implement and which to defer.

User research in Phase 1 surfaced valuable insights for advanced data usage tracking, such as state and local data, web metrics like Google Analytics, and highly localized geographic and demographic data, but many of these elements were not feasible to implement within the scope of the prototype. These remain better suited for future iterations. Despite the constraints, regular coordination and documentation ensured that user needs were translated into feasible features, prioritized, and that the prototype was delivered on time.

Implications for Phase 3 and beyond

Some features we determined were important based on user feedback, but we decided to defer them for future consideration. These included visualizations on publication type, Google Analytics integration, researcher contact tools, and AI-driven dashboard recommendations. Key gaps remain in capturing qualitative feedback about the DUP, addressing state and local data usage needs, and expanding usage analytics. Some agencies expressed interest in replicating DUP functionality internally; however, the longer-term vision is for a consolidated platform rather than multiple look-alikes. This highlights the need for continued engagement with agencies to clarify which features are most valuable and to determine whether new modules or extensions could meet those needs within the DUP itself. Community forums would be more valuable with active agency participation, and metadata scoring tools could improve data quality.

On the more technical end of the spectrum, the use of LLMs for tagging large volumes of data has room for improvement as cheaper and more powerful models become available as demonstrated by the smaller-scale LLM-as-a-judge automated QA module. In addition, more creative ways to approximate the usage of data assets in news outlets could be further explored to address the limited availability of full-text news sources.

These enhancements warrant further feasibility and usability testing before integrating into the broader NSDS ecosystem.

Phase 3: Communication and Documentation

Goals and objectives

Phase 3 supported the ability of a range of stakeholders to understand, adopt, and extend the DUP prototype and its findings. Building on the research and development work of Phases 1 and 2, this phase focused on creating a strategy to support and communicate how the DUP can serve as a shared service to inform public trust and engagement with a future NSDS.

We created a targeted communication strategy, produced tailored products for federal and non-federal audiences, and documented the technical, design, and user research outputs to support reuse and scaling. Collectively, this phase laid the foundation for continued adoption, funding, and refinement of the DUP as a sustainable part of the federal open data infrastructure.

Communication strategy and products

Communication plan

The success of the DUP prototype and its future adoption relies on key stakeholders being aware of the platform and understanding how they can use it for their own purposes. We collaborated with NCSES to develop a robust, targeted external communication plan that would clearly articulate the value and raise awareness of the DUP prototype among multiple stakeholder groups. This plan describes what information should be communicated, with whom, in which formats, and through which channels.

The communication plan is designed to equip NCSES to strategically engage those key audiences in both the short and long term in ways that align with the goals of the broader NSDS. In the short term, the activities described in the external communication plan will help NCSES conduct regular outreach to increase awareness of the DUP prototype among federal agencies. In the longer term, the activities described in the plan will help NCSES increase federal agencies' understanding of the DUP as a shared service and increase adoption of the DUP across other audiences that use federal data.

Audiences

Communication products and planned activities focus on engaging a broad range of data users and data creators, as identified through user engagement activities in Phases 1 and 2.

The primary audience for outreach efforts in the short term is federal agency staff. By focusing on reaching this group initially, NCSES can engage those who may be interested in adopting the DUP as a shared service under the NSDS umbrella and connecting their agency's data assets to support a potential future platform.

Additional secondary audiences that may help achieve longer-term goals of increased understanding and adoption of the DUP include (1) federal agency staff who may be interested in using the DUP for data access and understanding, (2) federal policymakers, (3) data journalists, and (4) academic and policy researchers. (Refer to the "Stakeholder engagement overview" section of this report for a more detailed description of each group.)

User engagement activities from Phases 1 and 2 revealed some differences between federal and non-federal audiences in their main interests behind using the DUP. Federal audiences were largely interested in understanding how their data are used—who is using them, for what purposes, and in what contexts. Non-federal audiences were largely interested in being able to easily navigate the fragmented landscape of federal data sources and identify popular and often-cited data sets and emerging data topics.

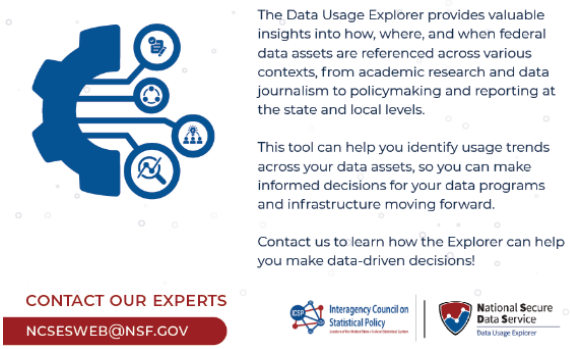
Content planning and generation

We collaborated with NCSES to develop clear, engaging, and accessible dissemination materials and messages. We tailored these products to effectively communicate the DUP's value proposition to the primary and secondary audiences and address their pain points, as informed by the user feedback sessions from earlier phases of this project.

The materials created to help NCSES implement the communication plan use NSDS branding to support a consistent, cohesive, and recognizable identity:

- **Dissemination kit.** This kit contains messaging and resources that NCSES and its partners can use to communicate about the DUP with key audiences. Components of the kit include (1) talking points to use when engaging with different audiences verbally or in written communications; (2) two decks of customizable PowerPoint slides and presenter talking points for internal and external presentations about the DUP, as well as a slide template; (3) a collection of social media post templates that focus on value proposition statements of the DUP, designed to be shared across America's DataHub Consortium (ADC), NCSES, and the social media channels of other partners; and (4) informative blurbs that can be included in newsletters and other email communications to describe the DUP and the challenges it can help address.

- **Postcard.** This postcard communicates high-level information about the DUP and guides people on how and where interested audiences can learn more about it. One version of this postcard features messaging intended to appeal to decision makers at federal agencies and is focused on adopting the DUP as a shared service. A second version of this postcard is geared toward helping individual agency staff advocate to their leadership about using the platform to access and understand data. Both versions of the card also contain contact information and a placeholder for a QR code that readers can scan to learn more. Each postcard can be used as a leave-behind resource at conferences or can be shared digitally as a PDF file following smaller meetings.



Outreach

As stated in the external communication plan, a variety of communication channels can help NCSES reach key audiences, amplify the DUP prototype, and promote future adoption and refinement. Channels include those owned by NCSES and new and existing partners, which can increase visibility and bolster the DUP's credibility. For example, websites and listservs—such as those for the NSDS, NCSES, ADC, and SAP—can be used to share information and build awareness about the DUP with primary and

secondary audiences via web copy, newsletters, and email outreach. Social media accounts—such as X, Facebook, and LinkedIn—managed by NCSES and other federal agencies can be effective ways to communicate with secondary audiences that may be interested in learning more about the DUP. Relevant data and research conferences—such as the Association for Institutional Research Forum or the Association of Public Data Users Annual Conference—can serve as venues to directly reach more specific audiences, such as early-career researchers. Finally, new and existing partners can serve as trusted intermediaries to further extend and amplify the reach of DUP-related communications beyond NCSES' direct network.

Technical documentation

Clear, accessible documentation was a core deliverable of the prototype to ensure transparency and enable future adoption. The project team produced a full set of online reference materials using Sphinx, an open-source documentation platform widely used across the data and software community.

The documentation explains the DUP prototype's architecture and data ingestion pipeline in plain language, while also providing automatically generated reference materials drawn directly from the codebase. This dual approach means that readers, whether technical staff, decision makers, or new collaborators, can explore both conceptual overviews and detailed technical references in one place.

The materials are designed for flexibility: they can be viewed in a standard web browser, updated alongside code changes, and extended easily as the platform evolves. With live preview features, contributors can see updates in real time, supporting collaborative development and ongoing refinement.

Lessons learned

Implications for Phase 3 and beyond

Outreach and communication. User feedback sessions provided early insights that were invaluable for establishing which messages and products to develop to communicate the value propositions of the DUP to different audiences. Any future short- or long-term efforts to increase awareness of and engagement with the DUP among primary and secondary audiences should continue to draw on these user feedback sessions. Opportunities to collect additional new feedback on DUP messaging and communication products can also help better gauge their effectiveness and allow for informed adjustments and the development of additional products, if needed.

Technical documentation. We used Sphinx to create clear, web-based documentation that combines plain-language guidance with auto-generated technical references. This approach improved transparency, streamlined collaboration, and ensures the DUP prototype can be more easily understood, updated, and adopted in the future. A future state of the DUP should continue to follow this approach to ensure that documentation stays up to date.

Discussion

Summary of recommendations

NSDS integration

The DUP could serve as a natural entry point into the NSDS, particularly within its Discovery and Navigation function. By providing transparent insights into how federal data assets are used, the DUP may orient users as they begin their evidence-building journey. Its integration with other NSDS services, such as the Data Concierge or Communities of Practice, presents opportunities for the platform to evolve into a central hub for both novice and expert users.

AI integration

AI offers promising ways for enhancing usability and reducing burden on the government. Tools such as chatbots, automated QA, and personalized feedback features may strengthen the platform's responsiveness. Continued exploration and refinement of these methods as they become more powerful at reduced costs—paired with careful testing and user feedback—can help clarify their role in future iterations of the DUP and NSDS.

User engagement and communication

Ongoing engagement remains essential to ensure the platform adapts alongside user needs and the broader federal data environment. Building on the communication products developed in Phase 3, there are opportunities to refine messaging, broaden outreach methods, and continue learning from diverse audiences, including journalists, researchers, and state and local data users. Over time, this sustained engagement can help build awareness, trust, and adoption across multiple communities.

Data coverage and expansion

Findings from Phase 1 and 2 highlighted strong interest from state and local organizations in understanding how their data are used. Expanding the DUP's coverage beyond federal sources, alongside integration of richer usage metrics, may broaden its relevance and utility across multiple levels of government and research communities.

Governance and sustainability

As the platform grows, clear approaches to metadata standards, QA, and community features will be important for maintaining transparency and trust. There may also be opportunities to explore governance models that promote stewardship of the DUP by agencies, helping ensure long-term sustainability.

Conclusion

The DUP prototype demonstrates the feasibility and value of a shared service that makes federal data usage more transparent, navigable, and actionable. By curating data asset references across academic, policy, media, and report domains, and by presenting them in a user-centered dashboard, the DUP responds directly to needs identified through extensive user engagement. Its modular architecture,

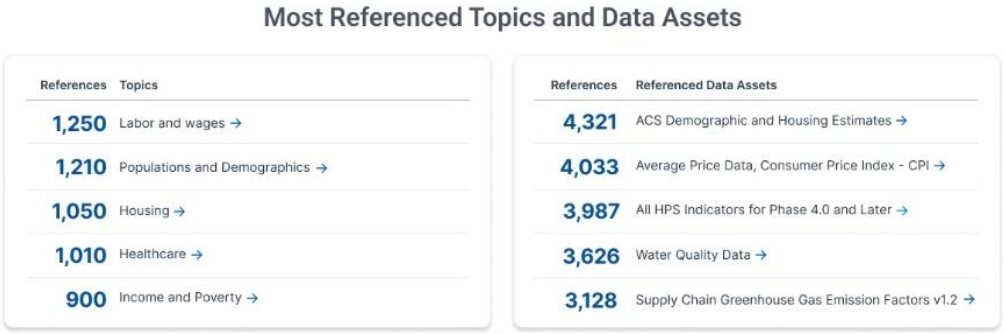
integration of AI-enabled QA, and open-source design choices provide a strong foundation for scaling and adaptation within the broader NSDS vision.

The project also underscored the importance of collaboration across agencies, academia, and the broader data user community in shaping requirements and testing design choices. Future success will depend not only on broader adoption but also on agency involvement and stewardship, ensuring that the DUP reflects diverse perspectives, remains sustainable, and continues to build public trust. By championing the platform, agencies can help establish the DUP as a cornerstone of the federal open data ecosystem.

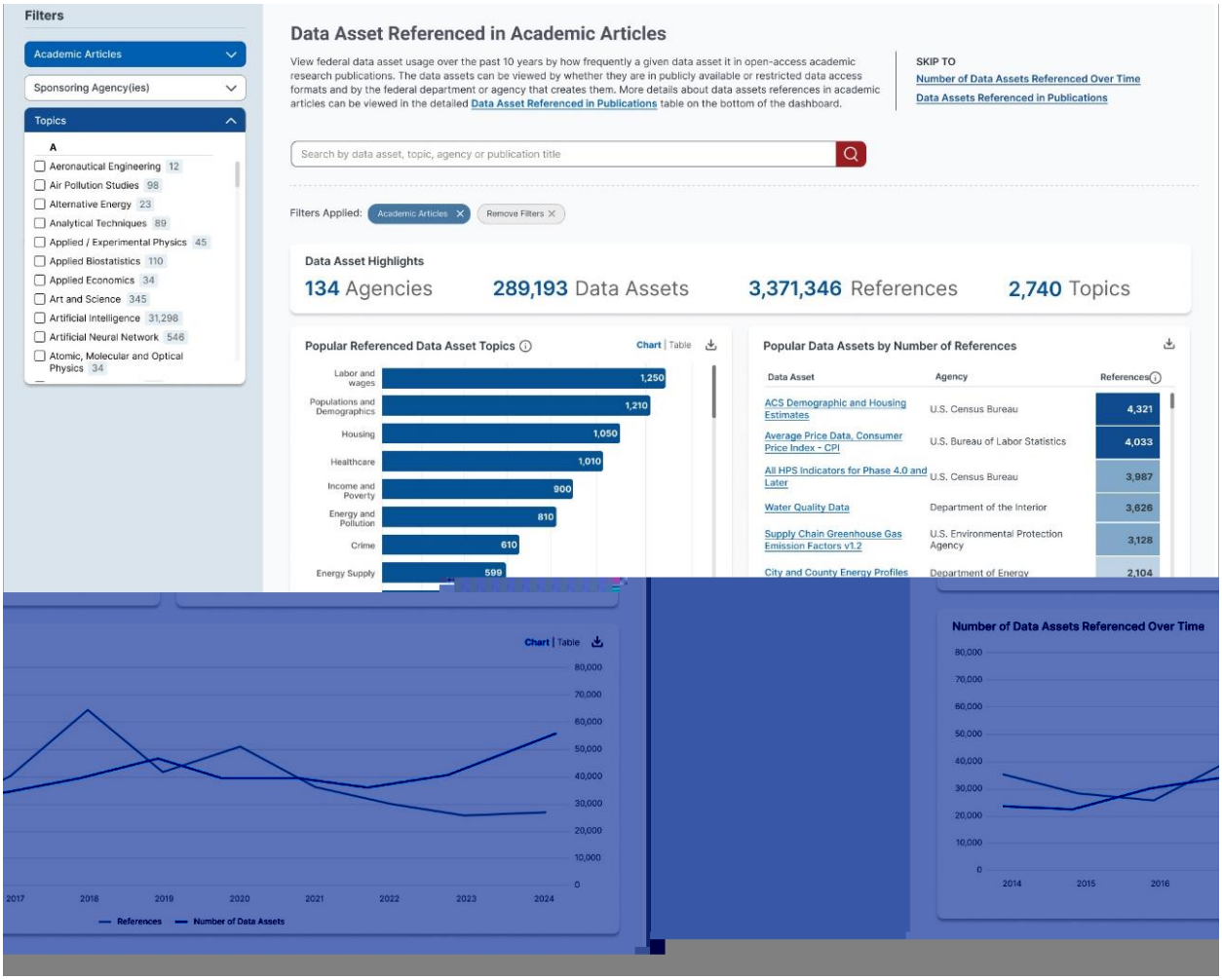
Looking ahead, the DUP's evolution presents opportunities to expand coverage to state and local data, strengthen metadata standards, refine AI integration, and enhance community features. With continued engagement, clear governance, and a commitment to transparency, the DUP can become a durable, shared resource that supports evidence building, fosters trust, and anchors navigation across the growing NSDS ecosystem.

Appendix A. Final Mock-Up of Key Components

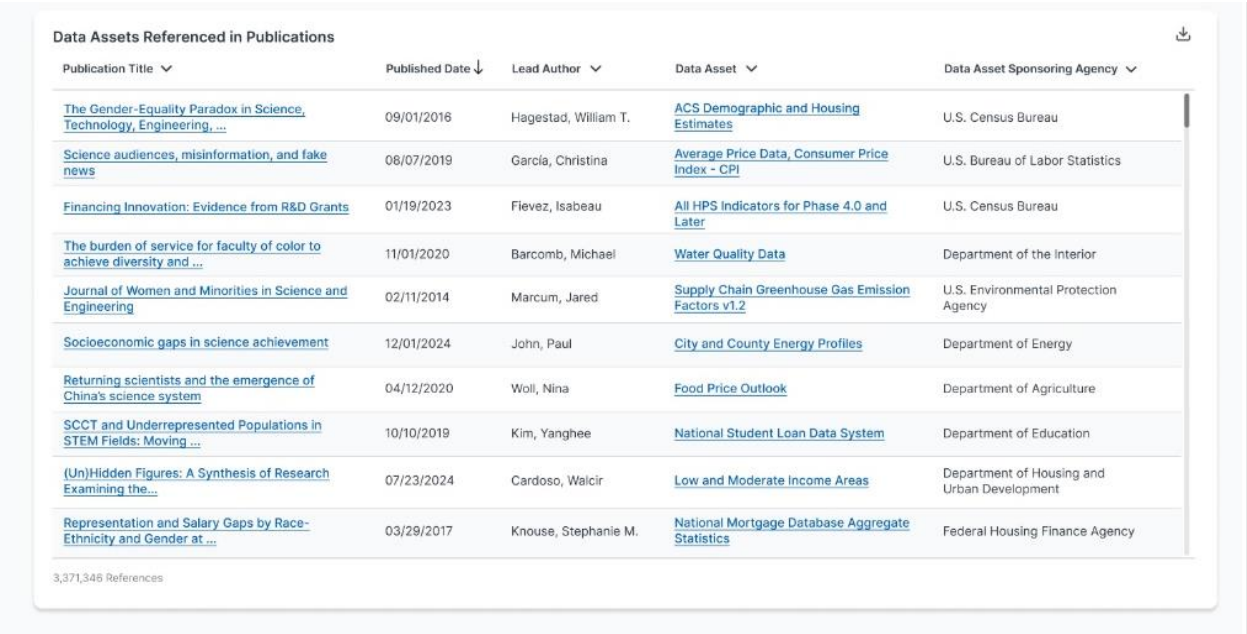
A.1. Home page



A.2. Dashboard data visualizations



A.3. Data assets referenced in publications

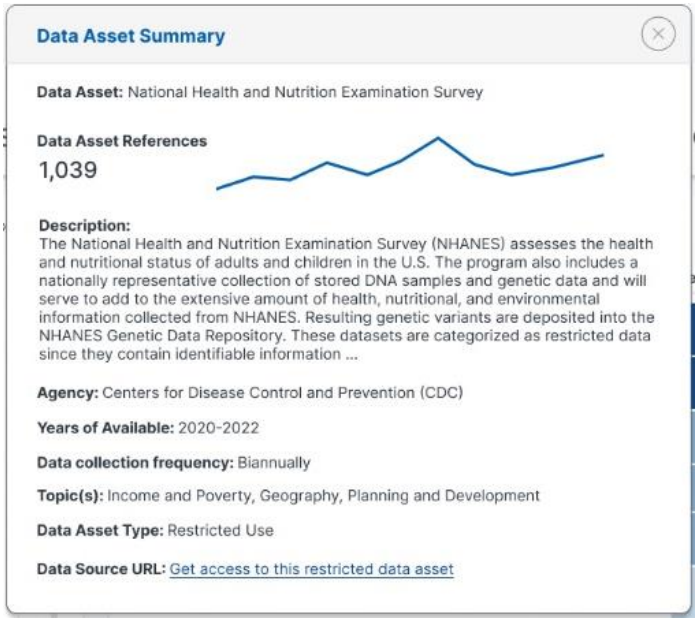


Data Assets Referenced in Publications

Publication Title	Published Date	Lead Author	Data Asset	Data Asset Sponsoring Agency
The Gender-Equality Paradox in Science, Technology, Engineering, ...	09/01/2016	Hagestad, William T.	ACS Demographic and Housing Estimates	U.S. Census Bureau
Science audiences, misinformation, and fake news	08/07/2019	García, Christina	Average Price Data, Consumer Price Index - CPI	U.S. Bureau of Labor Statistics
Financing Innovation: Evidence from R&D Grants	01/19/2023	Fievez, Isabeau	All HPS Indicators for Phase 4.0 and Later	U.S. Census Bureau
The burden of service for faculty of color to achieve diversity and ...	11/01/2020	Barcomb, Michael	Water Quality Data	Department of the Interior
Journal of Women and Minorities in Science and Engineering	02/11/2014	Marcum, Jared	Supply Chain Greenhouse Gas Emission Factors v1.2	U.S. Environmental Protection Agency
Socioeconomic gaps in science achievement	12/01/2024	John, Paul	City and County Energy Profiles	Department of Energy
Returning scientists and the emergence of China's science system	04/12/2020	Woll, Nina	Food Price Outlook	Department of Agriculture
SCCT and Underrepresented Populations in STEM Fields: Moving ...	10/10/2019	Kim, Yanghee	National Student Loan Data System	Department of Education
(Un)Hidden Figures: A Synthesis of Research Examining the...	07/23/2024	Cardoso, Walcir	Low and Moderate Income Areas	Department of Housing and Urban Development
Representation and Salary Gaps by Race-Ethnicity and Gender at ...	03/29/2017	Knouse, Stephanie M.	National Mortgage Database Aggregate Statistics	Federal Housing Finance Agency

3,371,346 References

A.4. Expanded metadata and pop-ups



Data Asset Summary

Data Asset: National Health and Nutrition Examination Survey

Data Asset References

1,039

Description:
The National Health and Nutrition Examination Survey (NHANES) assesses the health and nutritional status of adults and children in the U.S. The program also includes a nationally representative collection of stored DNA samples and genetic data and will serve to add to the extensive amount of health, nutritional, and environmental information collected from NHANES. Resulting genetic variants are deposited into the NHANES Genetic Data Repository. These datasets are categorized as restricted data since they contain identifiable information ...

Agency: Centers for Disease Control and Prevention (CDC)

Years of Available: 2020-2022

Data collection frequency: Biannually

Topic(s): Income and Poverty, Geography, Planning and Development

Data Asset Type: Restricted Use

Data Source URL: Get access to this restricted data asset

A.5. Community forum and data usage feedback module