New Approaches to Facilitate Responsible Access to Sensitive Urban Data

Andrew S. Gordon, Rebecca Rosen, Daniel Castellani, Daniela Hochfellner, Julia Lane

OVERVIEW

Improving government programs requires analysis of government administrative data. Providing access to these data to academic and public sector researchers is an important first step to robust analysis. At the same time, these data contain Personally Identifiable Information (PII) and so great care must be taken in **obtaining, storing, and providing access to these data**.

The Data Facility at **NYU's Center for Urban Science and Progress (CUSP)** is building on a long history of research on how to facilitate data curation, ingestion, storage, and controlled access in a safe and trustworthy environment.

The CUSP Data Facility combines computer science, information science, and social science approaches that include:

- building a **data model** that accommodates sharing research data across disciplines,
- employing **data curation and ingestion services** so that data providers can confidently share their data with authorized researchers,
- **providing secure computing environments** to work with these data,
- **converting data restrictions** into concise, easy to understand, and searchable metadata to help researchers find appropriate data for their research,
- **capturing activity around datasets** as contextual metadata so researchers can discover new data to complement their analyses.

DATA MODEL FOR SHARING

Comparing datasets across domains can be difficult. Focusing on a broad (rather than deep) and inclusive data standard can help researchers evaluate datasets that might fall outside of their domain of expertise

- Based on the Data Catalog Vocabulary (DCAT) data model for broad sharing
- Incorporates ICPSR subject thesaurus
- Includes detailed policy information for users to assess accessibility
- Facilitates metadata elements across different domains

DATA CURATION

By knowing the community you are serving, you can provide value on top of existing data sources by linking, standardizing formats for common features like geospatial variables, and creating combinations of data that you find get requested and used often

- Selection and provisioning of data in ways that are meaningful to the community you serve
- Combination of regularly used datasets into more dynamic data services
- Gives rise to reusability

SECURE COMPUTING

Data restrictions may preclude sharing data widely on the Internet. CUSP employs computing infrastructure that allows users to access sensitive data, while still complying with access restrictions set on those data

- Ability to share and collaborate between institutions and remotely in environments not directly accessible from the Internet
- Defined by user roles, careful management of group memberships, and who may access which datasets
- Infrastructure for FedRamp and Title 13 compliant computing in the cloud

SIMPLE DATA RESTRICTIONS

Data restrictions can be highly contextual. Creating overarching classifications for data and applying them consistently can cut down on time spent understanding when and how a dataset may be used

Public Data
- Public and freely accessible
- Public but providers would like to employ access controls or limit who may use

Restricted Data
- De-identified data with usage restrictions
- May contain sensitive data, but no personally identifiable information (PII)

Microdata
- Data which contains PII
- Located in secure, air-gapped environment
- Can make de-identified derivative datasets

CAPTURE USER ACTIVITY AROUND DATA

Description of a dataset alone is not enough to determine whether it is useful for a project. Capturing and displaying how it has been used by others in research, analysis, and publications can draw new insight into its value and its relationship to other datasets

- Seeks to understand how researchers utilize datasets across research projects
- Annotation and recommendation features to help users enrich and find datasets that match their work

GOING FORWARD

Through the creation and operation of a functioning data facility for researchers to access and analyze sensitive urban-centered data in new ways, we have made important observations and have a greater understanding of what succeeds. This knowledge will nourish future directions of this effort

- We have learned that users can adapt to complicated policies around sharing and access if you make it easier for them to do so
- We have demonstrated that an educational environment is a good place to help users understand new ways of interacting with data
- We now know that needs and expectations of users should drive the design of discovery and access infrastructure
- We anticipate that building data access and discovery around user needs will increase trust and usage of the data facility

NYU | Center for Urban Science + Progress