ACDEB
Advisory Committee on Data for Evidence Building

Focus Area 3: Technical Infrastructure

Amy O'Hara and David Park, ACDEB Members

May 2021

Overview of Proposed Technical Infrastructure Scope

- **People** – What are NSDS users expected to know, will tools be available to meet analysts where they are? What technology do we want for Access and Identity Management (AIM)? How will users authenticate? Will systems remember users, would users need to get another full background check if they'd been "sworn for life" already? What technology monitoring can keep users "safe," can monitoring use discern between a confused user and an insider threat?
- **Findable Data** – Catalog is part of Evidence Act, Metadata development is part of the FDS. Accessible – will technology apply legal and regulatory terms of use for automated access? Interoperable – if there's a will to share data, NSDS needs to have the way to do it. Need infrastructure that gets data from owner to compute, links files. Reusable – revisit what the commission said, meant, and wanted. If NSDS is not keeping extracts, how to promote reuse? What to archive, for how long? What are IP terms for researchers using NSDS? Technology must handle version control, provenance, expungement/sealing.
- **Safe Projects** – NSDS needs adequate data discovery and metadata, NSDS needs functional single application portal. Need transparency on requests, approved projects, and findings.
- **Safe Settings** – Technical infrastructure depends on (1) where you sit – enclave, specific site, or unspecified, (2) how you access data – IT in enclave, VDI, laptop/device, need AIM to align. How to federate access across existing secure platforms? How to have automated scaling of resources?
- **Safe Outputs** – Approaches for descriptive stats, coefficients, tables, microdata sample? Compare Statistical Disclosure Limitation - traditional, simpler methods vs Newer, harder to apply/explain methods. Protect inputs, protect outputs? Do users understand this? Will NSDS just dictate what happens or will it be user-input? Who makes decisions on tradeoffs? What do privacy budgets look like over time (some data do not age well). Don't focus on specific solutions like Synthetic Data or Secure Multiparty Computation (protects inputs but you still need to deal with the outputs). These all depend on threat models that need to be discussed.

ACDEB
Advisory Committee on Data for Evidence Building

Initial Low-Hanging Fruit

Potential Quick Wins

- Make list of sources/uses that prohibit virtual access
- Develop and apply standards
- Prioritize (a) federal agency data, (b) single agency asks, (c) joins between agencies with established linkage protocols
- Invest in automated labeling for terminals, logs and output, and disclosure packages
- Hire and train people who will clear the hurdles/create vaults where new technology can be used
- Get reciprocity agreements for FEDRAMPed systems
- Develop standardized process to review and safelist new software

ACDEB
Advisory Committee on Data for Evidence Building

Committee Discussion

Discussion Questions

- Should we start with federal agency held data, and focus on the easy joins first—records with SSN, EIN, ICD, NAICS, etc.? Then support the development of standards and incorporate state and local data, sensor data, private sector data, other data (unstructured, audio, video).

- How can we support linkages of data that the feds don't "own" or are decentralized/unstandardized? Court and criminal justice, EPA, food security, homelessness, worker credentials, building permits, etc. How will that data get to NSDS? Can't expect APIs at source agencies anytime soon. Should data be "shaped up" before it goes to NSDS or will NSDS have tools/labor to manage?

- What has to be in a physical enclave and why? Must it have the FSRDC lab/SCIF requirements? Brick and mortar is not viable option unless NSDS starts paying bills given full cost model of an FSRDC.

- How would a National Research Cloud fit in to address current memory, storage, and software constraints in agencies and FSRDC network?

- What are threat models? Who decides on priorities? What can reduce risks of an analyst doing something bad or careless?

ACDEB
Advisory Committee on Data for Evidence Building