11-2018

Technology for Civic Data Integration

Natalie Evans Harris
Bright Hive

Amy Hawn Nelson
University of Pennsylvania

DOI https://doi.org/10.23889/ijpds.v3i5.1042
Report: https://metrolabnetwork.org/wp-content/uploads/2018/09/Technology-for-Civic-Data-Integration.pdf

Technology for Civic Data Integration

Abstract
Efforts to collect, manage, transform, and integrate data across administrative systems into actionable knowledge to inform better policy decisions are becoming more common. However, the technical processes, procedures, and infrastructure they employ vary substantially. Variety in approaching data infrastructure, transfer, linking, and security is expected in this emerging field, but both established and developing efforts would benefit from cohesive guidance regarding the technical considerations of data integration, with focus on presenting a range of options that can be weighted based on context specific restrictions (e.g. cost, staffing, or existing infrastructure).

Actionable Intelligence for Social Policy (AISP), MetroLab Network, and the National Neighborhood Indicators Partnership (NNIP) with support from the Annie E. Casey Foundation, are convening a working group to shape and develop guidance on information architecture and technical approaches for data integration efforts such as those in the AISP and NNIP networks and the AISP Learning Community. This guidance will help newly emerging efforts as well as established ones looking to update their current approach. It will also inform policymakers and researchers who need a primer to better understand the technical components and considerations at play for data sharing and integration. This presentation will present findings, best practices and recommendations from this brief that will be released in Fall 2018.

Comments
DOI https://doi.org/10.23889/ijpds.v3i5.1042

Report: https://metrolabnetwork.org/wp-content/uploads/2018/09/Technology-for-Civic-Data-Integration.pdf

THE ANNIE E. CASEY FOUNDATION

Technology for Civic Data Integration

Amy Hawn Nelson, AISP

NNIP NATIONAL NEIGHBORHOOD INDICATORS PARTNERSHIP

ACTIONABLE INTELLIGENCE FOR SOCIAL POLICY

METROLAB NETWORK

Uncover And Navigate The Key Factors Influencing Successful Adoption Of Integrated Data Systems

This paper serves as:

- A **resource** to help agencies measure integrated data capacity
- A **primer** to better understand the technical components and considerations at play for data sharing and integration
- An **examination of factors** that influence the adoption of technology for data integration

This paper is not:

- The **answer** to all integrated data infrastructure questions
- A **vendor recommendation** list

Contributors

Lead Author: Natalie Evans Harris

Expert interviews, from different sectors and levels of experience with data integration

Joy Bonaguro, City and County of San Francisco
Matt Gee, Brighthive
Lisa Green, Data Domino Lab
David Hill, UNC Charlotte Urban Institute
Bill Howe, University of Washington
Anjum Khurshid, University of Texas at Austin
Julia Koschinsky, Center for Spatial Data Science
Jason Lally, City and County of San Francisco
Graham MacDonald, Urban Institute
Christopher Mader, University of Miami
Kathy Pettit, National Neighborhood Indicators Partnership
Deepthi Puram, University of Miami
Matthew Tamayo-Rios, Open Lattice
Emily Wiegand, Chapin Hall
Bill Yock, Santa Clara County

NNIP NATIONAL NEIGHBORHOOD INDICATORS PARTNERSHIP

ACTIONABLE INTELLIGENCE FOR SOCIAL POLICY

METROLAB NETWORK

What Have We Learned

Today's Landscape

- **People** are most important part of infrastructure (users, IT support, data scientists, etc)
- Industry is moving away from monolithic to **modular enterprise systems**

Key Technical Factors

- Security/privacy
- Transparency/Interoperability
- Identity and Reconciliation
- Mapping data use to methodology

Pain Points

Lack of Agility

- **Dependence on vendor** to run inquiries - Procurement processes should require data be interoperable and accessible
- Support linking new streams

Lack of Standards

- Lack of **data descriptions** the further up the data lifecycle

Lack of Government Alignment

- Shared **procurement** to reduce costs and increase interoperability
- Improved **governance** to streamline legal and political challenges
- Reduce **duplication** of efforts

Technology for Civic Data Integration

The purpose of this report is to **describe key considerations in building and sustaining IDS** and the various technology approaches that may be helpful in overcoming challenges in data integration.

Consideration 0: Staffing Expertise

Consideration 1: Data Management

Consideration 2: Security and Privacy

Consideration 3: Data Collection

Consideration 4: Data Storage

Consideration 5: Data Linking

Consideration 6: Data Access and Dissemination

A digital version of this report can be viewed here: http://bit.ly/2xaQW2L

Consideration 0: **Staffing Expertise**

Staff with **substantive business expertise** regarding the issues, programs, and context-specific social service areas is critical.

Infrastructure and technology choices **tightly integrated with the underlying work culture** and business processes.

Four key skill areas:

Data Storage and System Administration: solution architects, security administrators, database administrators, etc.

Data Integration: data / information architects, data modelers, data engineers, integration/interface engineers, etc.

Data Analytics: research / evaluation specialists, data analysts, data scientists, business intelligence analysts, etc.

Data Publication: website administrators, UX designers, BI portal administrators.

Key Question: ***Are the technology solutions chosen on-premise, cloud-based, or a combination? Are existing staff able to support both on-premise or cloud-based systems, or are new staff needed?***

Consideration 1: **Data Management**

Evaluating data management needs requires:

- Close examination of how to **drive standardization** upstream and downstream
- Plan for an **agile and flexible data intake and normalization** or indexing process

Reduces:

- dependence on specific vendors
- ongoing maintenance and upgrade costs

Supports ability to link new data streams as they become relevant

Key Questions: ***Does the solution support monitoring of data collection, cleaning, and integration activities? How does data quality feedback get passed to the data generators, and how does one track reporting deadlines and data completeness?***

Consideration 2: **Security and Privacy**

“How will the IDS protect the data?”

- **Individual privacy** must be legally, technically, procedurally, and physically maintained throughout the process.
- Technology leveraging **anonymization techniques** such as data masking, data aggregation, and data obfuscation.
- Process for ensuring **proper consent** to use administrative data.
- **Strong encryption algorithm(s)** are deployed (e.g. AES, DES, RSA, & ECC) and that the system administrator can manage the encryption keys which keep the data private.

Key Question: ***How does the solution manage user accounts to ensure authorized access to the correct data, at the right times, for the right reasons? Is there a single sign-on solution? Are there role-based access controls (RBAC)?***

Consideration 3: Data Collection

An IDS must be **dynamic and agile** enough to support new requirements, changing schemas, and new data streams.

Ways to upload data:

- **Manual** - an individual uploads the data
- **Automatic** - a connection between source and integration system
- **Hybrid** - some manual uploads and some automated connections

Key Question: ***What is the cost of establishing the connection and implementing technology necessary to automatically collect and clean the data?***

Consideration 4: **Data Storage**

How will the data stored - on-site, in a data center, or leverage the many cloud-based solutions

Where will the data be stored - Warehousing solution is needed to properly store and make data accessible

- **Repository** - The simplest warehousing options store the source data
- **Data lake** - Due to the increasing complexity of data sources, many organizations have moved to data lakes, which are a system of repositories.

Key Question: ***If using a cloud storage provider, is it up-to-date on data center and industry certifications such as HIPAA, FedRamp, PCI DSS, SSAE 16? What happens if there is a data breach?***

Consideration 5: **Data Linking**

Data linking is the process of **integrating different data sources**, based upon common business keys and other identifying information

There are two key methods for data linkage:

- **Deterministic matching** - considered more precise since it looks for exact matches in the content and format of datasets (i.e., identical SSN)
- **Probabilistic matching** (also known as fuzzy matching) - looks for closeness in the data (i.e., identical SSN with or without dashes) and provides weighted scores for likelihood of matching.

Key Question: ***What happens when a new data source is introduced and the matching rules need to change?***

Consideration 6: **Data Access and Dissemination**

Secure data access for analysts includes VPN remote access, limited and licensed data sets, and on-site access.

Data dissemination is the finished product and comes in many forms, from reports, to machine-readable datasets, to dashboards and websites.

- how the technology manages access to the finished product,
- how it will be disseminated, and
- how risks of redisclosure are minimized.

Key Question: ***Who gets access to the data infrastructure, and by what means? Are there levels of access for identified and de-identified information, or for internal and external users?***

A Look into the Crystal Ball

In Five Years:

Staff skills should move up in the value chain, especially as the technologies evolve. Technology will facilitate people to provide higher value deliverables; **how we integrate data will no longer be the question**, but rather how we ensure that the analysis and models are affecting practice and driving change.

Infrastructure should provide more than just integration support, it **should provide transparency and documentation about how solutions were developed**, including confidence levels for results.

The field has come a long way in the past five years, particularly around transparency and consistency. In order to conduct multi-site inquiry and continue to improve data models, all sites must emphasize the **development of metadata**; not only as a data operation but to contribute to the field.

The data science “gap” will diminish as schools incorporate analytic methods and skills into more undergraduate and professional graduate programs. **An ongoing bottleneck will be legal and transactional issues** such as privacy, domain use, ethics, international considerations, and inclusive engagement.

Experts and practitioners will be thinking about blockchain, consent management and identity management in **how we collect data on people**. We will need to be mindful of artificial intelligence and machine learning to ensure that we are careful in training data sets while focusing on ethical use and data governance practices.

THE ANNIE E. CASEY FOUNDATION

Thank you.

Amy Hawn Nelson, ahnelson@upenn.edu

NNIP NATIONAL NEIGHBORHOOD INDICATORS PARTNERSHIP

ACTIONABLE INTELLIGENCE FOR SOCIAL POLICY

METROLAB NETWORK