# The Organizational Structure of the Nationwide Data Initiative: Considerations

Robert L. Grossman
University of Chicago and Open Commons Consortium

## Introduction

In this note, we address some of the organizational issues that may to be faced by the Nationwide Data Initiative.

The Nationwide Data Initiative is designed to support programs and a related data infrastructure so that cities and states can more easily make use of evidence when planning and implementing programs and policies. Large scale administrative data from the Census Bureau and other federal agencies holds the promise of playing a critical role for evidence-based planning, but bringing this data together, linking it, and analyzing it is beyond the means of most cities and states.

## Background

In this note, we distinguish between:

- the *Program*, which selects and funds research programs and research projects;
- the *Projects*, either internal or external, that are selected for support; and
- the *Commons*, which are built by the Projects, and perhaps, by the Program.

In general, Projects are state and local governments and their partners with a plan to build a Commons containing administrative records and related data and to analyze the data to develop the evidence required to develop policies, put in place procedures, share derived data, and take other actions to achieve desired the objectives and outcomes.

We use the term *commons* or *data commons* when data, storage and computing infrastructure, and commonly used software services, tools and applications for analyzing data are co-located and managed as a resource for an evidence-based research community.[1]

An example of a commons is the National Cancer Institute Genomic Data Commons (GDC), which manages, analyzes and shares genomic and associated clinical data for NCI-funded researchers. The GDC is the largest, or one of the largest data commons in the world.[2]

It is helpful in the discussion below to distinguish between:

- the *Sponsor* or sponsors who fund the Program and the Commons and
- the *Manager* who manages the Program and Commons.

There may be separate Program and Commons Managers, with the Commons managed via a subcontract to the Program. The Sponsor is also sometimes called the Funder, especially when the organization is less involved in the day to day operations of the Program and/or the Commons.

Because of the difference roles and the different legal agreements required, it is also helpful to distinguish between:

- *Data Contributor Agreements*, which are required when organizations or researchers contribute datasets to the Commons;
- *Data Use Agreements*, which are required when researchers remove data or derived data from the Commons; and
- *Data Service Agreements*, which are required when researchers use the Commons to explore or analyze data.

Of course, a given organization or researcher may sign each of these if they are involved in each of these roles. Data Contributor and Data Use Agreements are standard whenever data is collected or generated by one party and used by another party. On the other hand, Data Service Agreements are relatively new and occur when a data platform such as a Commons is provided as a resource to the research community.

## The Analytic Diamond Model

In this section, we introduce (per Figure 1) a simple framework that may be useful when discussing the Program, the Projects they select for support, and the Commons that the Projects will develop and operate.[3] The foundation for the commons is secure, compliant cloud-based infrastructure for managing, analyzing and sharing large datasets (*analytic infrastructure*). Once the data are imported, cleaned, and linked, they can be explored, analyzed and statistical models can be built to study the problem or problems of interest (*analytic modeling*). Sometimes the hardest part is deciding how to use the models to achieve the desired outcome: e.g., determining a new policy, developing data products or scores that can be shared, taking certain actions such as intervention (*analytic operations*). Deciding what problems to tackle, how to tackle them, and how to measure success is part of *analytic strategy*. Coordinating these activities, which usually span an organization is part of *analytic governance*. Finally, generally project or program governance is needed that spans all the relevant stakeholders including, the Program, the community or customers it serves, the data providers, the sponsors, etc.

## Initial Focus of the Program

There are three separate dimensions that may be relevant when choosing initial projects:

1. Are the projects sufficiently compelling that the value provided by the data commons is compelling? In other words, could the analyses be done elsewhere with an infrastructure that was easier to build.
2. Is the evidence provided by specific analyses useful to problems that are widely acknowledged to be useful?
3. Is the value of the commons and the analysis obvious to the Sponsor?
4. Can some analysis be early in the project showing the value of the approach to the primary stakeholders?

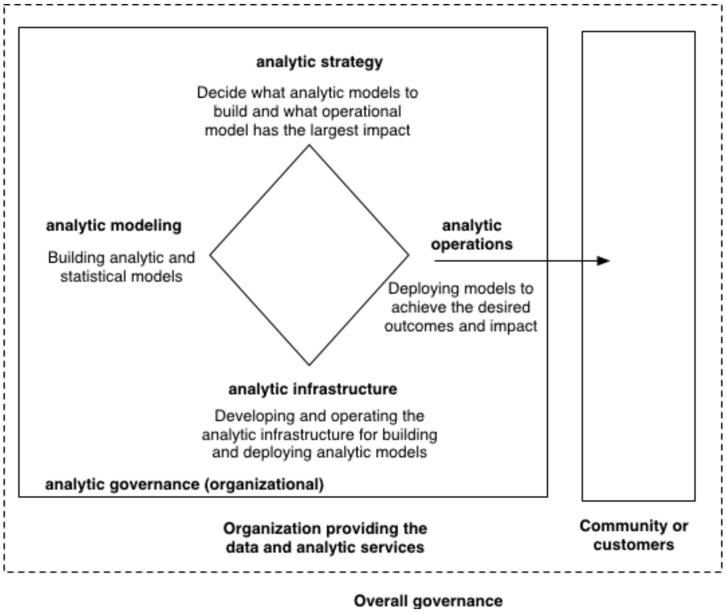

analytic strategy
Decide what analytic models to build and what operational model has the largest impact
analytic modeling
Building analytic and statistical models
analytic operations
Deploying models to achieve the desired outcomes and impact
analytic infrastructure
Developing and operating the analytic infrastructure for building and deploying analytic models
analytic governance (organizational)
Organization providing the data and analytic services
Community or customers
Overall governance

Figure 1: A model of analytic processes

It can be helpful to divide the datasets in a commons into three different types:

- **Core Primary Datasets (CPD).** Usually when building a data commons, there are some large third-party datasets that must be imported, integrated and linked, and analyzed. For example, with the Nationwide Data Initiative these would be datasets from the Census Bureau, the Social Security Administration, and the Internal Revenue Service. This usually requires substantial effort and time, perhaps as much as a year or more. Once imported, these datasets become a core resource for the research community.
- **Core-derived Datasets (CDD).** Once various third-party core datasets are imported and linked, it is often possible to analyze them and produce derived and aggregate datasets that can drive and accelerate research. If these datasets are aggregated sufficiently, they can often be open access, which of course greatly simplifies their use by the research community.
- **Project-specific datasets (PD).** There are other datasets associated with a project, such as a particular city or state analysis.

## Incentives for Data Contributors

The two main benefits to those that contribute data to the Commons are:

- Access to the other core primary datasets and core derived datasets and the ability to use these to accelerate their own research.
- Access to a powerful computing environment with all the associated software services, tools and applications that is also secure and compliant.

## Governance and Organizational Structure

Prior to discussing the governance structure, it is helpful to consider different options for the organizational structure:

- **University based.** One option is to base the initiative at a University. The advantage of this structure is that universities host many similar types of projects. As an example, the NCI Genomic Data Commons is hosted at the University of Chicago.
- **Independent not-for-profit.** Another option is to the base the initiative at an independent not-for-profit. An important advantage of this structure is the flexibility provided. Examples are the public-private BloodPAC Consortium and the Open Commons Consortium, an independent not-for-profit focused on developing and operating data commons and cloud computing to support the research community.[4]
- **Hybrid model.** A hybrid model combines a University presence and a 501(c)(3). The OCC-NOAA Data Commons is an example of this model, in which the Open Commons Consortium and U.S. National Oceanographic and Atmospheric Administration signed a Cooperative Research and Development Agreement (CRADA) to build an environmental data commons. However, it is unfunded. The OCC is responsible for raising the required funding. The University of Chicago provides essential in-kind re- sources and operating support to operate the Data Commons, but as a OCC Member, versus the operating entity itself. This makes it much easier for other Universities to get involved.

Generally, a commons has several committees and boards as part of its governance structure:

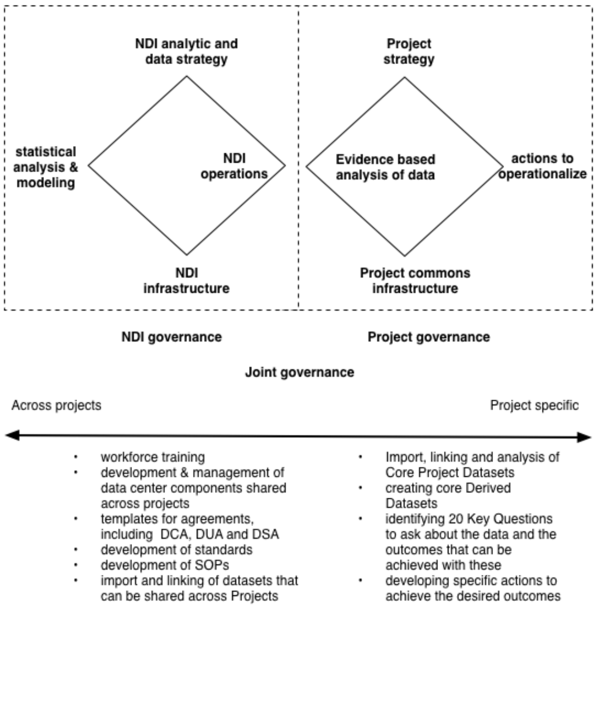
- **Executive Committee.** This committee is the core group that runs the commons and the associated program.
- **Scientific Advisory Board.** This board provides general scientific advice and guidance on the general direction, research initiatives, projects selected, etc. Once the program is up and operational, it is often helpful to set up additional committees.
- **Technical Advisory Board.** This board provides general technical advice and guidance on the development, operations and technology of the data commons.
- **Data Committee.** This committee, which can include members from the research community, decides which data get is accepted to the commons.
- **User Committee.** This committee, composed of external third-party users, provides advice about the operations of the program and commons from the users' perspective.
- **Advisory Council.** This group helps raise funds for the program and increase its visibility. It can have a variety of names, including Advisory Board, Advisory Committee, Board of Visitors, etc.

We close this section by highlighting the importance of addressing liability when considering the organizational structure of the Program and the Commons. It is prudent to assume that all Commons, no matter how well they operated will suffer one or more breaches, and to set up and operate the organizational structure for the Program and the Commons with this in mind.[5]

## Supporting Projects

Figure 2 shows one way of viewing the interaction of the Program with the Projects selected for funding. Many of the challenges faced by the different Projects will be common. One approach is for the Program to develop services and components that can be shared across projects, including workforce training, the development of standards for sharing administrative data, and key common services necessary for building and operating data commons.

Data commons, and more generally, analytic projects often fail. A simple mechanism to reduce the chance of failure is called Ten Questions.[6] With this approach, very early in the project a set of questions are developed that can potentially be answered from the data, and, if answered, can indicate the potential to quantify outcomes of societal value. The initial focus is ensuring the analytic infrastructure, the analytic operations, the analytic modeling, and the analytic governance are sufficient to answer the ten questions. The questions may be as simple as: "analyze the data to provide the evidence for Option A versus Option B for an upcoming policy" to "analyze the data and develop a score that can be updated each month and gives the likelihood that a student will drop out of school within the next 90 days."


NDI analytic and data strategy
statistical analysis & modeling
NDI operations
NDI infrastructure
Project strategy
Evidence based analysis of data
actions to operationalize
Project commons infrastructure
NDI governance
Project governance
Joint governance
Across projects
Project specific

- workforce training
- development & management of data center components shared across projects
- templates for agreements, including DCA, DUA and DSA
- development of standards
- development of SOPs
- import and linking of datasets that can be shared across Projects

- Import, linking and analysis of Core Project Datasets
- creating core Derived Datasets
- identifying 20 Key Questions to ask about the data and the outcomes that can be achieved with these
- developing specific actions to achieve the desired outcomes

Figure 2: Program roles for supporting Projects

## Funding Structure

The aspects of a Project that need to be funded include:

1. core data commons administrative and key technical staff
2. building of the data commons, including any data portals, and analysis tools
3. ongoing enhancements, refactoring, and maintenance of the data commons
4. cloud services (either private, public or hybrid cloud)
5. specific projects, including importing, cleaning, and processing the data
6. users engaged in specific analysis of project data

The first three are fixed costs that will vary depending upon the requirements. The next three are variable costs and depend upon the number of projects, amount of data, number of users, and the complexity of the analysis.

*It is critical to find a sponsor to cover the bulk of the fixed cost items.* It is possible that the Cloud Service Providers may provide donations to support cloud service costs) in part. Project- and user-specific costs can be raised on a per project basis if required.

One funding model is for the Sponsor to cover the cost of setting up and operating the Commons and the cost of modest use of the Commons (exploring the data and small-scale computing), with larger computing costs paid by the researchers. Researchers can be given allocations for the computing they use from Sponsor, can use a credit card, or can obtain separate funding for this purchase.

## Time Line

The time line will be primarily determined by the effort required to import and integrate the Core Primary Datasets and the complexity of the data portal that will be developed. Questions determining a time line for building a commons include:

- Will a new data commons be developed, an existing open source data commons be customized, or an existing open source data commons be used "as is"?
- Will data simply be made available in a secure environment or will the core primary datasets be processed, integrated and linked?
- Will core derived datasets be produced?
- Will a custom data portal or other user interface be developed or will the standard interface that comes with the data commons be used?
- How much Quality Assurance (QA) and security and compliance testing will be done?

We have developed approximately nine data commons to date, with some taking three months, and some taking 24 months, depending how these questions are answered. On the other hand, if data are simply loaded into an existing data commons, then the data commons can be launched as soon as the data is up- loaded and quality checked.

The following steps are usually done when setting up a commons:

1. Set up a development, QA, staging, and production commons

2. Define the data model
3. Import and Quality Control (QC) the imported data
4. Process the core primary datasets and required project specific datasets and associated analysis
5. Develop the data portal and any required specialized software
6. QA
7. Launch

We generally favor developing a data model (Step 2) and then automating the importing and processing of the data using an API based upon the data model (Step 3). Alternatively, the data can be uploading as uninterpreted tables, analyzed and then imported. Step 5 can be done currently with Steps 2-4.

### Executive Director and Other Senior Executives

Recruiting an appropriate Executive Director will be key to the success of the Nationwide Data Initiative. The ad used to recruit the BloodPAC Executive Director is in Appendix A.

A successful Executive Director is likely to have the following skills:

- **Organizational Leadership.** Organizational responsibilities include:
  - Establishing an operational framework for:
    - Setting up and operating the Program
    - Setting up and operating the associated Commons
    - Tracking projects and initiatives
    - Pilot collaboration requirements
    - Regular review of the Program, Commons and its pilots, projects and imported data
  - Leading decision-making processes with the Executive Committee and Scientific Advisory Committee
  - Leadership in developing program, organizational and financial plans with the Board of Directors and staff and carrying out plans and policies authorized by the board
- **Administrative Responsibilities.** Administrative responsibilities include:
  - Ensure that the Program follows all required financial and administrative controls required
  - Ensure that Program are consistent with relevant federal, state and other regulations governing the operations of the Program, including those related to conflict of interest
  - Ensure that Commons has the policies, procedures and controls required to manage the operations of the commons in a secure and compliant fashion
- **Fund raising and Sustainability –** Drive short and long-term fund raising, and to develop and execute a sustainability strategy
- **Communications, Messaging and Public Relations. –** Develop or oversee the Program's branding and messaging and integrate the Program's communications so that donor outreach, publications, website and social media, marketing and media relations are in sync with, and support, a unified and effective message

Other Senior Executives should include a chief data officer, a chief technology or information officer, and a chief information security officer.

## Measures of Success

Perhaps the most important measure of a success is: were the ten questions answered, and, if so, what was the societal impact? From a Commons perspective, the major measures of success are usually taken to be:

1. How many users are analyzing the data in the Commons, and how many research papers are being written?
2. What are relevant metrics to measure the value of the research enabled by the Commons? What is the impact of this research as measured by these metrics?
3. How many projects are using the Commons?
4. How many projects are contributing data to the Commons?
5. How many sponsors have supported the Commons?
6. Less important measures, but still relevant:
    - Is the operating model being copied by others?
    - Is the technical model being copied by others?
    - Is the software used by the Commons open source, and, if so, what is the level of contribution and reuse of this software?

## Staffing

The following staffing is usually used to develop and operate a Commons:

- **Software developers.** Software developers are needed to develop, maintain, and enhance the Commons.
- **DevOps.** DevOps technical staff are needed to operate the cloud computing infrastructure required by the project. The number of DevOps staff will depend upon the whether a public, private, or hybrid cloud is used and the scale of the cloud.
- **Users Services.** Users services staff will be needed to answer user questions and to provide outreach.
- **Data scientists/analysts.** Data scientists/analysts will be needed to import, clean and integrate the data; and process and analyze the core primary datasets to create the core derived datasets and analyze the project specific datasets.
- **Project manager.** Generally, a full-time project manager is needed. If the project is large enough, additional project managers will also be needed.

The level of staffing will be driven primarily by:

- the complexity of the imported data;
- the number and complexity of the derived data created;
- the complexity of project specific data portals; and
- the number of users and the support that they require.

## Infrastructure

As mentioned above, the first choice regarding infrastructure is the nature of the cloud: public clouds, such as Amazon's AWS, Google's GCP, and Microsoft's Azure; private clouds; or hybrid clouds that integrate both public and private clouds. At small scale, public clouds probably make more sense. At intermediate-to-large scale (20+ racks), hybrid clouds probably make more sense.

If public or hybrid clouds are used, the second question is whether only a single public cloud will be used (e.g. AWS) or whether multiple public clouds will be used. Each cloud has a rich variety of services, and researchers sometimes prefer one over another. Initially, for small projects, starting with a single public cloud makes sense. As the scale of the project grows, multiple public clouds sometimes make sense. If multiple clouds will be used in the future, planning for this is important.

Another important question is whether a proprietary or open source architecture will be chosen will be used for the software stack and whether the commons will interoperate with other commons (in a safe and compliant fashion).

Standards for data commons are still emerging. The OCC may get involved in trying to facilitate standards. For certain communities, such as biomedical data, some organizations are trying to organize standards (e.g. GA4GH), but even in these cases, standards are still quite immature.

## Training and Workforce Development

Data science is understood to be the intersection of: mathematics/statistics, computer science, and a specific discipline, such as public policy. In our experience, the challenges in training researchers with data science skills are not out of the ordinary. Numerous reports about training researchers with data science skills are available. Training researchers with data science skills is much easier using a data cloud or data commons than without it, in our experience.

## Other Considerations

Other considerations worthy of mention include:

- **Intermediate datasets.** Given the potential wide impact, a valuable initial activity may be producing intermediate aggregated datasets that requires less stringent security and can be more widely disseminated.
- **Accounting and billing.** The operations of a successful commons, de- pends upon a good logging, accounting, and billing system. This is the case even though most developers do not believe accounting and billing are important.
- **Security.** Despite any security measures in place, it is a safe assumption that the Commons will be breached, so it is important to plan for this eventuality.

## Additional Information

For the Open Commons Consortium approach to building and operating data commons, see

Robert L. Grossman, The Open Commons Consortium framework for a data commons, *white-papers.rgrossman.com/occ-framework-18-v2.pdf*, 2018.

For an example of a large-scale data commons (NCI Genomic Data Commons), see Izumi V Hinkson, Tanja M Davidsen, Juli D Klemm, Anthony R Kerlavage, and Warren A Kibbe, A comprehensive infrastructure for big data in cancer research: Accelerating cancer research and precision medicine, *Frontiers in Cell and Developmental Biology*, 5, 2017.

For general considerations when sharing research data, see Michael W Carroll, Sharing research data and intellectual property law: a primer, *PLoS biology*, 13(8):e1002235, 2015.

## Appendix: Example of an Ad for An Executive Director

In 2016, the BloodPAC Project moved from the Moonshot Office at the White House to the independent not-for-profit Center for Computational Science Research Inc., the parent of the Open Commons Consortium (OCC). Below is the ad that was used to recruit the Executive Director for the BloodPAC Consortium.

### BloodPAC Executive Director

The duties of the BloodPAC Executive Director, include, but are not limited to:

### Organizational Leadership

- Establishing an operational framework for:
    - Tracking projects and initiatives;
    - Pilot collaboration requirements;
    - Project and data review.
- Leading decision-making processes with the Executive Committee and Scientific Advisory Committee.
- Leadership in developing program, organizational and financial plans with the Board of Directors and staff and carrying out plans and policies authorized by the board.
- Providing weekly reports on the organization's operations and activities to all consortium members. Organizing and managing member relationships in order to build relationships of confidence and trust with all Consortium members and external stakeholders.
- Attracting and stewarding new Consortium members and collaborators.
- Organizing and leading all discussions and meetings with co-chairs and members.
- Building a solid foundation for the structural future of BloodPAC 2017 and beyond.
- Managing advocacy and professional society affiliations as well as regulatory, standards, and funding agency relationships.

### Administrative

- Ensure that the BloodPAC Consortium follows the financial and administrative controls required by the CCSR.
- Ensure that BloodPAC Operations are consistent with IRS and other federal regulations governing 501(c)(3) corporations, including those related to conflict of interest.
- Maintain official records and documents, and ensuring compliance with CCSR, federal, state and local regulations.
- Maintaining accountability for fiscal management and long-term fiscal planning for the consortium.
- Developing and maintaining sound financial practices, including ethical use of funds and resources.

- Working with BloodPAC members and legal support to address and develop collaborative research agreements, MOU's, data sharing agreements, and IP issues as they arise.

**Fundraising and Sustainability**

- Working directly with the Executive Committee and Board of Directors to establish short and long-term fundraising goals.
- Developing and implementing a comprehensive holistic fundraising plan.
- Generating ideas for new BloodPAC pilot study funding.
- Endeavoring to raise private dollars to support pilot studies.
- Establishing and cultivating productive relationships with corporate funders, individual donors, and philanthropic funds.
- Reporting and stewardship of donors.

**Communications, Messaging and Public Relations**

- Developing BloodPAC's branding and messaging.
- Integrating all BloodPAC communications so that donor outreach, publications, website and social media, marketing and media relations are in sync with, and support, a unified and effective message.
- Primary point of contact for all media relations.
- Managing an annual calendar of events ranging from point of entry opportunities to major donor level events.
- Maintaining BloodPAC's presence in the public arena, within the funding community and the community at large, providing visibility for the consortium and members.
- Working to increase visibility, awareness and usage of the Blood PAC Data Commons through continued dialogue with all members and potential collaborators in order to maintain consistent regulatory interaction, and recurring engagement with external stakeholders.
- Interacting and advocating for the public face of BloodPAC through public appearances, relationship development and enhanced public relations.

[1] Robert L Grossman, Allison Heath, Mark Murphy, Maria Patterson, and Walt Wells, "A case for data commons: Toward data science as a service," *Computing in Science & Engineering*, 18(5):10–20, 2016.

[2] Robert L Grossman, Allison P Heath, Vincent Ferretti, Harold E Varmus, Douglas R Lowy, Warren A Kibbe, and Louis M Staudt, "Toward a shared vision for cancer genomic data," *New England Journal of Medicine*, 375(12):1109–1112, 2016.

[3] Robert L Grossman, "A framework for evaluating the analytic maturity of an organization," *International Journal of Information Management*, 38(1):45–51, 2018.

[4] RL Grossman, B Abel, S Angiuoli, JC Barrett, D Bassett, K Bramlett, GM Blumenthal, A Carlsson, R Cortese, J DiGiovanna, et al., "Collaborating to compete: Blood profiling atlas in cancer (bloodpac) consortium," *Clinical Pharmacology & Therapeutics*, 101(5):589–592, 2017.

[5] Robert L. Grossman, Robert Tedesco, and Walt Wells, "The legal environment when sharing research with a data commons," in draft, 2017.

[6] Robert L. Grossman, *The Strategy and Practice of Analytics*, O'Reilly, in draft.