DATA DISAGGREGATION PROJECT

METHODS OVERVIEW

Stephanie Peterson, Learning & Evaluation Officer
Nadege Souvenir, Senior Vice President of Operations & Learning

AUGUST 2021

Introduction

As a community foundation, the Saint Paul & Minnesota Foundation has an obligation to support the whole of our community. One of the most public ways that we demonstrate that support is through our grantmaking. To achieve our aspiration – to contribute to making Minnesota an equitable, just and vibrant place where all communities and people thrive – our grantmaking efforts also need to be equitable and just.

2020 led us to take a deeper look at our grantmaking practices. In doing so, we acknowledged that we have not historically examined the whole of our grantmaking. Up to this point, we primarily focused our analysis on foundation-directed, competitive grants. As a result, we could not easily provide data to paint a picture of all of the beneficiaries of our grantmaking. This carries implications for our internal strategies as well as for our external partners, who increasingly have turned to organizations like Candid to better understand grantmaking more broadly; if we do not understand the whole of our grantmaking, we cannot expect external organizations to have a complete picture of it either.

Gaining a complete picture of our grantmaking is essential to ensuring that the Foundation is doing its part to support a thriving and equitable ecosystem throughout the East Metro and the entire state of Minnesota. We know from our community partners that there is a disconnect between what they know from their lived experiences and what data systems capture in the aggregate. We also know that major decisions are often made based on the data captured in these systems.

In order to see the complete picture, we knew that we needed to disaggregate all of our grantmaking data from both foundation-directed and donor-advised grants. Accordingly, we defined the following set of primary learning questions:

- How does the average grant amount vary across beneficiary communities and grant types?
- What percentage of our grants are going to certain population groups and subject areas?
- How much are we granting to certain population groups or in certain subject areas?

This project is the beginning. As we better understand our data, we can use it to inform our theory of philanthropy and guide future strategies and engagement with donors and community.

Methodology

After narrowing our set of primary learning questions, we determined that most of the answers lie at the intersection of a few key data elements: primary population served by the grantee, the

subject category of the organization, geography and basic grant information. Below, we've given a high-level summary of how we gathered the information and turned it into valuable insights about our work.

While we followed an iterative process, it can be summarized into the following four phases:

ONE	TWO	THREE	FOUR
			
Extract and Unify Data Fields	Create Org. Level Data Set	Visualize and Explore	Document and Plan

Phase 1: Extract data and create crosswalks.

To begin, we extracted grant-specific data from our grants management system into a series of CSV files by grant type. In our case, the information collected across grant types is not fully standardized, leading to significant variation in what has been collected for each type (e.g., donor-advised grants vs. grants from open competitive rounds). Therefore, we first standardized our existing fields across all grant types and blended the files into one.

Once we combined the data into one file, we were able to assess what needed to be done to meet our two objectives of (a) understanding our grantmaking through disaggregated data, and (b) submitting data to Candid. From there, we identified the additional information that we needed and crosswalked our fields with the required Candid fields.

Phase 2: Create a dataset at the organization level with additional information.

Two of our biggest data analysis challenges presented themselves early on. The first is that the amount and type of data we collect vary widely across grant types. This is particularly difficult because the grant type with the least amount of information, donor-advised, makes up the largest proportion of our grants. The second challenge is that collecting and maintaining the accuracy of organization-specific information is significantly more difficult when repeated across multiple grants, which naturally occurs due to the many-to-one relationship of grants to organizations. Given these challenges, we decided to structure our dataset much like a CRM system, with two separate tables – one table of organization-specific information and one of grant-specific information – which get blended together as appropriate for analysis.

Based on the assessment of our data in Phase 1, along with our plan for analysis, we were able to identify the additional fields that we needed to fill in our dataset at the organizational level.

In our case, we needed fields for primary population served and subject area, along with a few supplemental fields that would help us make those determinations more efficiently, such as mission statement and URL. This created a shell of what essentially could be thought of as a big worksheet that had many empty cells to fill in.

Our next step was to fill in that worksheet, and we first turned to our existing internal information to make a first pass. Our system holds the most information on programs that have received competitive grants, so we applied that information at the organization level where appropriate. This includes information collected through grant applications, grant reports and grant officer knowledge.

We needed to collect the information for the remaining empty cells from outside sources. The most efficient way to do this was to start by leveraging a small selection of fields from publicly available IRS 990 data via a web scraping tool. The three most helpful things pulled were NTEE codes, mission statements and URLs.

- NTEE codes were primarily used to determine a broad subject area for the organization.
- Mission statements were helpful in determining primary populations served.
- The organization's URL address added a helpful layer of efficiency, since we would still need to look up many of these organizations manually to determine subject area or population served.

While the 990 data sped the process up considerably, we still completed a significant amount of work by hand to ensure we could feel reasonably confident in the information. 990 data isn't without its own set of challenges and still requires a human eye. Additionally, there are some types of organizations that either don't submit 990s or have incomplete or inaccurate data.

Phase 3: Visualize and explore!

Once we had both the grant-specific and organization-specific datasets in a place where they felt complete and accurate enough for our purposes, it was time to begin exploring the data. We blended the data back together in such a way that would allow for flexibility in analysis and used a data visualization tool that would facilitate a wide range of interactive options.

Phase 4: Document process and plan for the future.

While this project will continue to evolve and grow, we took the time to document our process and what we have learned about a number of things, including:

- Where our own internal data practices need to be strengthened.
- Assumptions baked into decisions made along the way.
- Practical reminders to ourselves for repeating this work in the future.
- Guidelines on appropriate use of the historical data set, specifically which questions it is designed to answer and which it is not.

We are also in the process of planning for the future! This includes plans not only for how future data collection will need to evolve across all grant types, but also how we evolve as an organization based on the trends we see. Projects such as these offer us a unique opportunity to reflect on our own theory of philanthropy.

Additional Context and Information

Subject area:

Subject area buckets enable us to answer a set of questions related to topic areas receiving funding, and the industry standard for classification is the NTEE Code. This multi-level taxonomy was developed by the IRS and the National Center for Charitable Statistics to classify nonprofits.

The challenge with any classification system is often not with the system itself, but with the varying interpretations of it by the people entering the data. For this project, whenever possible, we used as a starting place the NTEE category that the organization selected for themselves on their 990. Modifications were made where there was an obvious anomaly in the data or where the purpose of the grant was much more focused than the institution where the grant was made.

Populations served:

Identifying populations served is at the core of this data disaggregation project. While we initially focused on race and ethnicity, it became possible to expand our focus with the information that the 990 data provided. We selected a limited number of demographic and socioeconomic categories to begin our exploration based on the ease of classification and using the limited information available.

Identifying primary populations served was done through a multi-faceted and additive approach using the following methods.

1. Internal data that the Foundation regularly collects through grant applications, grant reports and program officer input.
2. Data available through the Index of MN Nonprofits Primarily Serving Communities of Color, compiled by the Minnesota Council of Nonprofits.
3. Mission statements analysis.
4. Manual look-up of organizations through their websites, social media pages, etc.

Learning Along the Way

Most foundation grants management systems are packed with data collected over many years, in various formats, and for various purposes that have changed over time. Wandering into this ocean of data can be overwhelming, but we learned a lot along the way. A few things seem worth mentioning:

- Take advantage of this time to do a data audit. Which fields are most reliable? Which fields need to be cleaned and standardized? Which fields are no longer used? Which fields need to be created? Where is institutionalized racism (and other forms of bias) baked into the data systems?
- Start by brainstorming the types of questions you would like to answer and the data that will be required to answer them early on. Having a prioritized list will help keep the project appropriately scoped throughout all phases. It will also help you to weigh the costs and benefits of the amount of manual work that should be done.
- Think big up front when determining population served categories. Building a historical dataset such as this is resource-intensive, and efficiency is key. Adding population categories later on will require repeating steps each time (e.g., another sweep of mission statements). Similarly, it is easiest to do multiple years at a time given the number of repeated organizations.

Partner and Tools

Given the large number of grants and the complexity of the data sets, we chose to utilize the expertise of our learning and evaluation partner (i3 Works) to help us build, manipulate and visualize the data.

Given the fluid nature of the project, they selected a set of tools that facilitated the iterative nature of our methodology and our desire to visualize our data in a flexible and interactive way. Specifically, the tools used were *Google Cloud DataPrep by Trifacta* for data transformation and *Google Data Studio* for visualization.

Conclusion

We provide this information with the hopes that it is helpful as you embark on your own data disaggregation efforts. We welcome conversation and the opportunity to learn from you as well. Please feel free to reach out to us at info@spmcf.org or by visiting spmcf.org.