Research article

Building a school data portal using open data: A description of the user research & development process

Steven Azeka a,b,*, Aankit Patelt, Kenjiro Kanamarub,d

a Communication, Media, & Learning Technologies Design, Teachers College, Columbia University, United States
b Robin Hood Learning + Technology Fund, United States
t STEM Education Programs, City University of New York, United States
d School of Management, Yale University, United States

ARTICLE INFO

Keywords:
School data
Comprehensive education plans
Data tool
Program officers
Computing education
Computer science education
Qualitative research

ABSTRACT

The lack of relevant and disaggregated data about schools is a challenge for organizations implementing and funders looking to invest in kindergarten through 12th-grade education programs in New York City (NYC), often forcing organizations to make critical programmatic and financial decisions with only partial or anecdotal information about schools and its existing efforts. Over the last two years, an institution of higher education and a private foundation worked together to address this problem by indexing text in thousands of annual school reports and combining multiple open data sets about schools, including geography, demographics, and academic performance, into a single database. The initial use case focused on understanding the existing computing education efforts in NYC. However, seeing the potential to inform a broader set of questions, the collaborators began exploring the utility of such data with program officers at the foundation. Interviews of program officers around their programmatic needs and challenges illuminated their struggles with analyzing scattered data, difficulty accessing non-academic data and the need for disaggregated and geographical data. These findings are being used to inform the development of a data portal that interfaces with the aggregated dataset for foundations and others interested in data about schools.

1. Introduction

New York City (NYC) is home to one of the largest repositories of open data in the United States, with over 2000 datasets published (The City of New York, 2021a) ranging from student attendance records to snowplow routes (The City of New York, 2021b). Education currently represents the most extensive open data category with over 1000 different datasets (The City of New York, 2021a) containing a wealth of information of the city’s nearly 1 million kindergarten through 12th grade (K-12) students and 1876 schools (The City of New York, 2021c). However, given the sheer size, fragmentation, and varied formats of these datasets, the layperson collecting information about schools, teachers, and students would struggle to find relevant data. The inaccessibility of this data is particularly challenging for organizations providing schools with funding, individualized resources, supports, and/or programming. As a result, many organizations obtain necessary programmatic information by individually interviewing or surveying the schools. The data are often inaccurate or limited (the person being interviewed or surveyed may not always have the latest information) and collected inefficiently (school-by-school basis). One source that could efficiently bring more accurate programmatic data at the school level is an under-utilized and largely inaccessible dataset called Comprehensive Education Plans (CEPs).

In NYC, school leadership teams of administrators, teachers, and parent coordinators work with their school community to plan for the upcoming school year, ultimately creating a CEP. These CEPs are published annually and are the only location containing comprehensive information on existing programming, curricula, academic interventions, and parent engagement. Gaining access to this data is particularly meaningful for those investing in the K-12 education space in NYC, as this is the only consistent location where schools identify education efforts. However, the NYC Department of Education publishes individual school CEPs as a Portable Document Format (PDF) that combines multiple documents, making large scale multi-school analyses challenging. A tool was needed for program officers investing in the K-12 education space.

In early 2020, a private fund, referred to moving forward as “the Fund”, began exploring what it would mean to address the inaccessibility

* Corresponding author.
E-mail address: sja2126@tc.columbia.edu (S. Azeka).

https://doi.org/10.1016/j.heliyon.2022.e10517
Received 27 November 2021; Received in revised form 28 March 2022; Accepted 26 August 2022
2405-8440/© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
of school data with the specific use case focused on computing education. The Fund supports several organizations that provide computing training and resources to K-6 teachers with the ultimate goal of helping young students build the foundations needed to be more open to experimenting with computers as they get older. One of the Fund’s largest grantees, the Computer Science for All (CS4All) initiative, set the ambitious citywide goal of providing an equitable Computer Science (CS) education to all K-12 students in NYC. Since its launch, CS4All has increased access to K-12 CS education with programs running at over 800 schools in NYC (New York City Department of Education, 2021) and supported by over 370 K-12 CS education organizations (Dvorkin, 2020). However, many NYC schools report its CS and broader STEM efforts in several locations, not all easily accessible by the public. The lack of information about existing efforts has contributed to an uneven distribution of CS teacher training and student opportunities (Dvorkin, 2020). Other organizations have recorded the location and general progress of CS implementations (e.g. Code.org, CSTA, ECEP Alliance, and CSforAll), however, the information primarily exists at the state level (Code.org, CSTA, & ECEP Alliance, 2021) or volunteered at the individual district and school levels (CSforAll, 2021).

In late 2020, a collaboration between the Fund and an institution of higher education was formed to explore if inaccessible but rich open data, like CEPs, may provide insights for organizations funding K-12 computing education programs in NYC. The institution of higher education was specifically interested in better understanding existing computing programs in K-12 schools with the goal of aligning its teacher education program. As the collaborators developed the school data portal, they found the collected data could inform a broader audience of program officers interested in other school-based initiatives (e.g. curriculum). To better understand how program officers might make use of this tool for their own analyses, the authors interviewed program officers at a private philanthropy regarding their current use of open data. The challenges surfaced during these interviews informed the future development of the school data portal. In this article, the collaborators outline the process of developing a school data portal, the research exploring the utility of a school data portal with program officers, ethics of working with school data, and future direction of the school data portal.

2. Development of the school data portal

The school data portal project began with analyzing a specific subset of schools’ CEPs to find if and where schools mentioned CS-related information. The collaborators initial exploration of CEPs led to creation of a web application to filter, group, sort, and aggregate schools based on information in their CEPs, such as a school’s vision for computing and technology education, the curriculum used, partner organizations, and participation in the CS4All initiative. The primary feature of this prototype was the ability for the user to search CEP texts. To make the text of CEPs searchable and to allow searches to be filtered by school year, school district, and community board district, data was retrieved from two sources: 1) a repository of CEPs hosted on the NYC Department of Education (DOE) URL nycenet.edu and 2) the Location Management database (LCGMS) available on the NYC DOE InfoHub.

The NYC Department of Education hosts all 14,879 CEPs (http://nycenet.edu/documents/oaisl/cep/) accessible by appending the school year, in the format of “YYYY-(YY-1)”, and borough and school number. All NYC schools are given a District Borough Number (DBN) in the format of “DDBNNN”, where D represents the school district, B the borough, and N the school number. Schools are identified in the CEP repository and LCGMS as the “BNNN” which is unique without the inclusion of the district number. The collaborators, using a school they knew had existed in NYC for decades began testing how many school years of CEPs were available and found that CEPs going back to the 2009-10 school year. The collaborators used a bash script (https://github.com/aankit/cep-access/blob/main/app/search.py) to download all CEP PDF files from the 2009-10 school year to the 2018-19 school year (last year available at the time) by checking all the potential combinations of school years, borough codes (B, M, Q, R, X), and numbers between 001 and 999.

A prototype of the school data portal used Pythonic utilities and frameworks to convert, store, index, and search the data. Individual CEP PDFs were uploaded to S3-compatible cloud storage (ceps.nyc3.digitaloceanspaces.com) and accessed by appending the school year and borough number (/2009-2010/K003.pdf). PDFMiner was used to extract the 915,361 pages of text elements. The URLs, text, and school information, including geographic and community board districts were loaded into a prototype SQLite3 database in a Flask environment using SQLAlchemy. The text was indexed using ElasticSearch and enabled as a database function using SQLAlchemy Declarative API. ElasticSearch was used to index text because SQLAlchemy does not support using relational database full-text search functions and could be implemented with a few functions that maintain a separation of concerns.

The Flask application currently exposes the data through a form. Users can utilize the form to search one or multiple terms and filter by school year, school district, community board district, or individual school name. The ElasticSearch query used to search the text can be viewed (https://github.com/aankit/cep-access/blob/main/app/search.py).

The resulting prototype (Figures 1 and 2) is a search tool allowing users to filter schools by geography, such as school district, community board district, and search term(s). A search for “computer science” using the tool shows the percentage of schools including the term in their CEPs each year and links to relevant pages in the PDF (Figure 3). While less than one percent of the nearly 1600 schools publishing CEPs referred to “computer science” between the 2009–10 and 2012–13 school years, that percentage increased to almost 13% by the 2018–19 school year. The change coincides with the launch of pilot CS teacher training in the 2013–14 school year that led to the 2015–16 launch of the CS4All initiative. The ability to see citywide CS trends is one demonstrated use for the tool.

As the collaborators began sharing the school data portal, they realized the prototype could benefit audiences beyond those in computing education. Program officers from other areas could use the school data portal to explore a wide range of school-based trends, like adoption rates of different curricula, the pervasiveness of socio-emotional and culturally relevant programs, or the level of services provided to special education students. Given the openness of search terms, the collaborators wanted to prioritize the most relevant and highest priority information for program officers.

![Database diagram](image_url)  
Figure 1. Database diagram.
3. User research of school data

During the summer of 2021, the collaborators formed a research team to investigate program officers’ existing usage of open data. The research team adopted a phenomenological methodology to capture the “universal essences” of a specific group’s lived experience for this study (Creswell and Clark, 2017). The “specific group” studied was a sample of K-12 program officers at a private philanthropy and their “lived experiences” involved in investing and developing different K-12 education programming. After conducting semi-structured interviews with the sample, the researcher (a team member from the Fund) used conventional content analysis to develop codes leading to themes around the phenomena inductively (Miles and Huberman, 1994).

3.1. Participants

A homogenous purposeful sampling method was used to identify five program officers and two managing directors at the philanthropy who actively invest in the K-12 education space. These individuals were co-workers of the researchers. The participants’ prior experience with computing programs ranged from peripheral integration into schools to deep involvement in creating computational thinking strategies. Participants who were not familiar with the K-12 educational landscape were excluded in the study. The research team decided to start with participants in philanthropy, given the number of grantees they manage and familiarity with the challenges the grantees face. Two participants self-identified as male, and five self-identified as female. The participants represented different races and ethnicities, with two self-identified as Black, four self-identified as White, and one self-identified as multi-racial. All participants consented to be interviewed and have deidentified results published.

3.2. Data sources

The research team conducted all participant interviews. The interviews were semi-structured and featured several key questions to help the researcher understand the challenges and affordances of open data. This structure allowed the researchers to ask follow-up questions. Interviews were approximately one hour long, recorded with the participant’s consent, and conducted online.

3.3. Data analysis

The research team transcribed all interviews and conducted data analysis using Dedoose, a cross-platform app for analyzing qualitative
text. The transcripts were coded through the process of horizontalization, whereby every statement was analyzed equally, and significant statements that provided insights into data challenges and data were coded and highlighted (Moustakas, 1994). The significant statements were used to generate larger clusters of meanings or codes. The researcher then reanalyzed all transcripts using the 53 codes that were established.

3.4. Results

Based on the data analysis, three overarching themes emerged amongst the seven participants: 1) challenges with analyzing scattered data; 2) difficulty accessing qualitative data; and 3) need for disaggregated and geographical data.

3.5. Theme #1: scattered data

The interviews revealed that participants found data analysis time-consuming as data is hard to gather. Across the board, participants shared a view about the importance of gathering school data in their work: “There’s a lot of data that you need to gather in order to really have a real sense of what’s happening [in schools].” (Participant #1, July 8, 2021). However, many participants expressed concerns over how, “there’s a tremendous amount of publicly available [school] data” and “gathering school data when it’s in a lot of different places can take a lot of time” (Participant #3, July 16). One interviewee described, “I would open eight different spreadsheets, I would copy the information… put them all in a new workbook… and it would take me a really long time” (Participant #5, August 3, 2021). Under these circumstances, participants find the steps of searching and gathering data necessary to analyze to be “time consuming and exhausting” (Participant #4, July 19, 2021). Overall, there were indications that “having an aggregate place to look [at school data]... would be really helpful” (Participant #1, July 8, 2021).

3.6. Theme #2: qualitative data

Participants also expressed high interest in qualitative data while identifying challenges gathering such data. All interviewees highlighted that quantitative academic data is essential, but qualitative data can add an extra layer of meaningful insight. One member indicated that “there’s so much that can impact a student’s academics, including things like social-emotional learning, and behavioral health” (Participant #1, July 8, 2021). Another interviewee raised the question of whether there is a way to “track how many schools have already used a social emotional learning curriculum” (Participant #4, July 19, 2021).

Despite the need for such qualitative school data, interviewees expressed challenges around searching and aggregating such data. One participant described the challenges in extracting qualitative data from the CEPs to determine the level of priority in computational thinking: “You [open] up every file finding the PDF... [and] you have to go in four steps down through the DOE website... it was like if you made me make a bowl of rice, and you told me I had to pick each grain of rice to put it into the bowl... it’s not that I can’t do it, [but] it’s just gonna take me hours” (Participant #5, August 3, 2021).

3.7. Theme #3: disaggregated data and geographic data

Participants expressed high needs for disaggregated data and geographical data that can help them make informed decisions. Disaggregated data can help decision-makers get essential insights and make meaningful comparisons for program monitoring and due diligence, especially when accounting for factors related to equity. An interviewee articulates the significance of taking one step deeper into the face value of school data: “Part of what [the foundation] wants to do is to serve the most vulnerable, and it’s really hard to do that if you don’t disaggregate across racial, socioeconomic, and then also other lines like students with individualized educational plans and English language learners.” (Participant #1, July 8, 2021).

Geographic data of specific student characteristics and schools is another need from the interviewees. One participant shared, “[For example] we’re interested in the English language learners... where the biggest concentration of English language learners [are].” (Participant #4, July 19, 2021). Understanding concentrations of students or schools with specific characteristics can help decision-makers “identify spaces with severe challenges” and help them set target populations to make focused investments (Participant #6, August 3, 2021). One interviewee shared a concrete use case, in which she has geographic data of “most likely eligible qualifying households with kids in the house under 17 who are not connected to public benefits that I call hot zones” and would like to see low performing community schools in these “hot zones” to reach out (Participant #7, August 5, 2021). Overall, “the hardest part is being able to organize [data] in such a way that it’s easy to draw analyses out” (Participant #5, August 3, 2021).

4. Future development of data tools

Based on the emergent themes from the interviews, the collaborators plan on revising the school data portal to include four critical features (Table 1). Features 1 and 2 will expand the existing school data portal to include student demographic and performance information, thus, consolidating several existing data sets into a single portal. Feature 3 further curates the school data portal to include qualitative data that is both accurate and informative. Feature 4 addresses the need to connect

Table 1. School data portal updates.

| Features | Interview Focus Areas | Backend | Design questions |
|----------|-----------------------|---------|-----------------|
| 1. Add school student demographic to database | Scattered data; Disaggregated and geographic data | Join NYC DOE demographic snapshot (e.g. gender, race, age, English language learners, special education, etc.) | - Are the demographic fields included relevant to users? - Is the way in which demographics are reported relevant/accessible to users? |
| 2. Add school performance data to database | Scattered data | School Performance Dashboard; State ELA & Math Test Scores; Graduation Rates | - Which variables of school performance are relevant? In what context are they being used? |
| 3. Add curriculum information for each school to database | Scattered data; Qualitative data | Utilize EdReports database of all existing curricula, not just those listed on website, to search CEPs | - Is the curriculum identified what schools are using? - Is the curriculum reported important to users? |
| 4. Zip code search | Disaggregated data and geographic data | Location Management Database | - Are users able to search for schools in the geographies they are most interested in using zip code? Would another geocoding variable be more useful? |
all prior data to a geographic region so school data portal users can disaggregate information by neighborhoods, districts, and boroughs.

The data school data portal will be revised using an agile methodology to meet the needs of program officers and other potential users, minimize negative consequences of the tool, and minimize the development time. The collaborators will engage users in discussions of datasets included, playtests of search and filter tools, and gather feedback on report structures during upfront design and planning sessions. They also plan to draw on other education focused data projects that have sought to bring together multiple sources of publicly data. These projects include national projects such as Education Data Portal and Ed Build as well as local projects such as the public data tools made available by the NYC Department of Education (https://tools.nycenet.edu/). Each of these data tools bring multiple sources of data about schools together for specific use cases. Education Data Portal aggregates federal and national organization, EdBuild aggregates and standardizes state funding data, and NYC Department of Education aggregates state and local data on schools. Notably the Education Data Portal and EdBuild include use cases where users can filter to find subsets of districts or schools with particular characteristics or can compare schools and districts. For example, users of the Education Data Portal can pull data on Title I schools in a specific locality and EdBuild offers comparisons of district funding per student relative to percentage in poverty. The NYC Department of Education tools do not offer the ability to filter for subsets and only offers limited comparisons between school performance.

5. Discussion

School data has the potential of playing a significant role for program officers’, and others who are invested in programmatic implementation, ability to better understand what is happening in schools and the potential impact of their efforts in schools. Through a series of interviews, the collaborators uncovered the challenges and larger needs of program officers. Based on the findings, the collaborators plan on revising the school data portal to give users: 1) access to additional school open data alongside the comprehensive education plans to address and not exacerbate the scattered data issue, 2) a summary of key attributes reported in comprehensive education plans useful to qualitative analyses, 3) view disaggregated data by school or subset of schools, and 4) the ability to search by granular or custom geographies. The collaborators envision the school data portal helping program officers in two crucial areas: finding appropriate schools for a given intervention and program monitoring.

Program officers often have difficulty recommending schools to grantees given the varying mission, student population, operational needs, and existing interventions. Building off our findings, the school data portal should allow users to enter the specific constraints and quickly identify all qualifying schools based on such factors as geography, existing programs, racial composition, poverty levels, or any other publicly available school-related data characteristics. For example, Participant #7 trying to target interventions at low-performing schools serving a specific population in a specific geography would benefit from being able to access robust qualitative data available in the database of CEPs by filtering on school performance, student population, and geography.

Program officers could use a host of existing open school data to monitor the progress of education efforts, however, the disaggregated nature of the data makes this challenging. One indicator demonstrating the prioritization of different education initiatives is the recognition of such efforts in the school’s CEP. Organizations can also access information around academic progress, demographic data, and student/teacher survey results. As mentioned above, these are often essential indicators to monitor equity. Gaining access to both will provide the first step to accessing aggregate data usually accessible by evaluators.

Over the next year, the collaborators plan on iterating on the existing tool to incorporate the insights from program officers outlined in this paper. The school data portal will be shared on GitHub and openly available. To ensure the design and function adequately meets the needs of program officers, the collaborators will conduct a focus group with key informants to preview the portal and gain further feedback. Further research and evaluation are needed to explore the utility of NYC open datasets and the specific effects of the school data portal.

6. Ethics of working with data collected from schools

The data used in the prototype tool and proposed data to enhance the prototype’s effectiveness are publicly available per city and state statutes. However, there are still ethical considerations when indexing publicly available documents about schools and joining these data with schools’ aggregate student demographics or performance data. Using the Federal General Services Administration Data Ethics Framework (General Services Administration, 2020) and considering the environment around school data and open data generally, the collaborators feel these recommendations for ethical data use apply to their project:

- Consider the impacts their data activities might have on the public, individuals, and communities and take measures to minimize any negative consequences.
- Document the data collection and curation process to promote an understanding of the data used in analysis and reporting and enable reproducibility of results.
- Allow for public access, amendment, and contestability to data findings, where appropriate and per organizational policies.
- Make data available for research equitably and objectively.
- Document data activities in ways that can be communicated to stakeholders.

The potential negative consequences of making open data more accessible include misrepresentations, biases, or omissions that result in disparate impacts due to stakeholder actions. For example, there is the potential for data to be lost, such as during conversion of PDFs to text or joining separately published datasets or the collaborators own biases regarding school data and schools to impact their school data portal design. While the collaborators heavily rely on school CEPs as sources of data on computer science curriculum and instruction, it is possible that schools are not reporting computer science instruction or other interventions because their superintendents do not prioritize it. Transparency of the limitation of the portal, how it was built, and providing stakeholders, including the schools the collaborators are reporting on, methods to contest or add to their data will help address these ethical issues. However, as described in this paper, the collaborators believe that proactively building a dialogue with their stakeholders to understand their needs, gather feedback, and test their portal’s affordances and constraints to identify issues could lead to negative consequences, especially for schools in underserved communities. This approach echoes an essential theme of engaging the public in the recommendations made in response to New York City’s Automated Decision System Taskforce (Richardson, 2019). While the school data portal will not be posed as an automated decision tool; the collaborators believe that the environment in New York City around the use of data concerning resource or service allocation warrants holding the school data portal to similar standards.

In addition to conducting the interviews described in this paper, the collaborators have engaged additional stakeholders by presenting their work at NYC Open Data Week. This presentation included information about the relevant state and federal data available and their purpose and asking attendees what questions they asked of the data they sought to make more accessible.

Declarations

Author contribution statement

Steven Azeka; Kenjiro Kanamaru: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.
Aankit Patel: Conceived and designed the experiments; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This work was supported by the Robin Hood Learning and Technology Fund.

Data availability statement

Data associated with this study has been deposited at Github: https://github.com/aankit/cep-access.

Declaration of interest's statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

Acknowledgements

The authors would like to thank all the interview participants in this project.

References

Code.org, CSTA, & ECEP Alliance, 2021. 2021 State of computer science education: accelerating action through advocacy. Retrieved from https://advocacy.code.org/stateofcs.

Creswell, John W., Clark, Vicki L. Plano, 2017. Designing and Conducting Mixed Methods Research. Sage publications.

CSforALL, 2021. CSforALL K-12 school and district pledge. Retrieved. CSforALL from: https://www.csforall.org/projects_and_programs/k-12-school-district-pledge/.

(Accessed 14 November 2021).

Dvorkin, Eli., 2020. Plugging in: Building NYC’s Tech Education & Training Ecosystem. Center for an Urban Future and Tech: NYC.

General Services Administration, 2020. Federal Data Strategy - Data Ethics Framework. https://resources.data.gov/assets/documents/fds-data-ethics-framework.pdf.

Miles, Matthew B., Huberman, A. Michael., 1994. Qualitative Data Analysis: an Expanded Sourcebook. Sage publications.

Moustakas, C., 1994. Phenomenological Research Methods. Sage publications.

New York City Department of Education, 2021. CS4ALL Program Update: Year 5 Progress & Recent Developments. https://tinyurl.com/aewdd9zz.

Richardson, Rachida. 2019. Confronting Black Box: A Shadow Report of the New York City Automated Decision System Task Force. AI Now Institute. Retrieved October 6, 2021, from http://ainowinstitute.org/ads-shadowreport-2019.html.

The City of New York, 2021a. NYC Open Data. NYC Open Data. Retrieved October 6, 2021, from https://opendata.cityofnewyork.us/.

The City of New York, 2021b. Introduction. Open Data for All: 2021 Progress Report. Retrieved October 6, 2021, from https://moda-nyc.github.io/2021-Open-Data-Repository/report/dataset-highlights/.

The City of New York, 2021c. DOE Data at a Glance. NYC Department of Education. Retrieved October 6, 2021, from https://www.schools.nyc.gov/about-us/reports/doe-data-at-a-glance.