Stories from the Frontier

Breakthroughs, Challenges, and Recommendations from the First Five Years of Open 990 Data

APRIL 2022

Dorothy A. Johnson Center FOR PHILANTHROPY

GRAND VALLEY STATE UNIVERSITY

COMMISSIONED BY

PHILANTHROPY & SOCIAL INNOVATION

aspen institute

Prepared For

The Aspen Institute's
Program on Philanthropy & Social Innovation

Washington, DC

Project Team

Jeff Williams, Director, CDRL
Trish Resurreccion Abalo, Research Associate

Dorothy A. Johnson Center for Philanthropy at Grand Valley State University

Grand Rapids, MI

Suggested Citation

Williams, J. & Abalo, T. R. (2022, April). *Stories from the frontier: Breakthroughs, challenges, and recommendations from the first 5 years of Open 990 data.* A collaboration of the Dorothy A. Johnson Center for Philanthropy at Grand Valley State University and the Program on Philanthropy & Social Innovation of the Aspen Institute.

Acknowledgments

The Johnson Center and Aspen Institute would like to thank the following individuals who shared their expertise and time as part of this effort:

- All of the people that offered their comments and reflections in an interview with the project team, including:
    - Chris Taggart and Rebecca Lee, OpenCorporates
    - Jacob Fenton, Journalist
    - Jacob Harold, Candid
    - James Sheehan and Charlotte McCary, New York State Department of Law
    - Jesse Lecy, Arizona State University
    - Jon Durnford, DataLake
    - Magda Guillen Swanson and Vicky Kelberer, Vanguard Charitable
    - Nathan Dietz, University of Maryland
    - Sheila Krumholz and Anna Massoglia, OpenSecrets
    - Woodrow Rosenbaum, GivingTuesday
- Serena Lai, Victoria Miller, and Krizia Pascuccio, William Randolph Hearst Fellows at the Aspen Institute, for their assistance in gathering and vetting the use cases
- Members of the Aspen Institute's Nonprofit Open Data Collective for their multiple suggestions of resources for the appendix and use cases, as well as encouragement and review of the early conclusions
- The Bill & Melinda Gates Foundation for their support

Dorothy A. Johnson Center for Philanthropy

The Dorothy A. Johnson Center for Philanthropy at Grand Valley State University was established in 1992 with support from the W.K. Kellogg Foundation. Our mission is to be a global leader in helping individuals and organizations understand, strengthen, and advance philanthropy, resulting in a smart, adaptive sector that helps create strong, inclusive communities.

We put research to work with and for professionals across the country and the world. Through professional education offerings; research, evaluation, and consulting services; and bold thinking to advance the field, we support a philanthropic ecosystem defined by effective philanthropy, strong nonprofits, and informed community change.

201 Front Ave SW, Suite 200
Grand Rapids, MI 49504
616-331-7585 // jcp@gvsu.edu

johnsoncenter.org

Dorothy A. Johnson Center
FOR PHILANTHROPY

The Aspen Institute's Program on Philanthropy & Social Innovation

Grantmaking foundations, nonprofit organizations, social enterprises and public-private partnerships offer lasting solutions to societal challenges. They are at the heart of civil society. The Program on Philanthropy and Social Innovation (PSI) seeks to inform and maximize the impact of these social sector actors through leadership development initiatives, convenings, and communications so that each can contribute to the good society at home and abroad.

PSI's Nonprofit Data Project strengthens the work of nonprofits on the ground, enhances transparency, and provides critical information and knowledge to those who work for and with nonprofits such as donors, government officials, members of the public, charity regulators, and scholars.

PSI is deeply grateful to the Bill & Melinda Gates Foundation for its support of this report.

For information on PSI's Nonprofit Data Project, contact Cinthia Schuman Ottinger at cinthia.schuman@aspeninstitute.org.

The Aspen Institute
2300 N St. NW, Suite 700
Washington, DC 20037

https://www.aspeninstitute.org/programs/program-on-philanthropy-and-social-innovation-psi/

www.aspeninstitute.org

PHILANTHROPY & SOCIAL INNOVATION
aspen institute

Table of Contents

Introduction 5

History of Open 990 5

What Open 990 has Meant to the Field 7

Challenges 9

Ease of use is a real concern. 10

Perennial gaps and challenges remain. 10

Inspiration and lessons learned from other Big Data projects 12

Recommendations 14

Collaboration is essential to create a coordinated Open 990 data ecosystem. 14

The IRS can take essential steps to improve the utility, consistency, and accessibility of the existing Open 990 data. 15

Provide support for the IRS tax-exempt staff — especially the data processing team. 15

Require federal agencies to disaggregate nonprofit data. 16

Encourage or require existing Open 990 data providers to be more transparent. 16

Conclusion 16

Appendix: Tools for Accessing the Open 990 Data 17

Where to access the Open 990 raw data 17

How to access and use Open 990 raw data 17

Examples of sites that use and leverage Open 990 data 17

Other sources of information about the nonprofit sector 19

Introduction

Open data projects have been in existence for decades, especially as the amount of data stored on computers throughout the world has skyrocketed. Accessibility to that data is at the heart of these efforts, as public and private entities work to make data freely available and useful to the public. Also critical is the role that freely available data in general — and public or government data in particular — play in accountability and transparency in government, as well as increasing both public participation and public awareness. As one interviewee noted, "Data makes it clear that the earth rotates around the sun — not the sun around the earth. Data can lay plain the places where our worldview needs to change."

The Open 990 Project of the Aspen Institute and its partners represents a giant leap forward, providing nonprofits a connected, data-informed future. After only five years, there are compelling examples available from individuals, nonprofits, and collaboratives alike of how the Open 990 Project is seeding and empowering change throughout the nonprofit sector. A large number of websites, projects, researchers, governments, and companies are now using IRS Forms 990, 990-EZ, and 990-PF data (hereafter, "990 data") to redesign how they work and how they engage with stakeholders.

The Dorothy A. Johnson Center for Philanthropy (Johnson Center) at Grand Valley State University (GVSU), in partnership with the Philanthropy & Social Innovation team at The Aspen Institute, is pleased to share:

- curated examples of current use cases from the Open 990 Project,
- a review of structural considerations for open data projects,
- inspirational and cautionary tales from other open data projects, and
- recommendations for ensuring a vibrant and collaborative future for the Open 990 Project.

Data makes it clear that the earth rotates around the sun — not the sun around the earth. Data can lay plain the places where our worldview needs to change.

History of Open 990

Prior to 2016, the only source of raw data containing IRS Form 990 information was the IRS itself in the form of multiple CDs/DVDs released on a monthly basis and sold by the IRS for thousands of dollars. Because the raw data was provided in image files, only the largest organizations such as the Urban Institute's National Center for Charitable Statistics (NCCS), the Foundation Center, or GuideStar had the technical teams in place to convert the image files into machine-readable databases and spreadsheets at a cost of close to $2 million[1] each year. Because of the time and computing power needed to process the images, even the largest data organizations were selective in choosing the size of nonprofits and the portions of the IRS Form 990 to regularly convert into usable formats for researchers, regulators, and the public. For example, for many years NCCS provided the 990 Core files which contained selected and cleaned financial and governance information from the 990 forms.

[1] Estimated from: Simone Noveck, B., & Goroff, D. L. (2013, January). *Information for Impact: Liberating Nonprofit Sector Data*, p. 18. https://www.aspeninstitute.org/publications/information-impact-liberating-nonprofit-sector-data/

Data were available, but not for every nonprofit, nor for every field contained on the 990 forms.

Building on the work of many pioneers in the field[2] — particularly the Urban Institute and NCCS, which applied steady pressure to promote e-filing of IRS Form 990s with both state and federal regulators, including developing e-filing software for nonprofits — the Aspen Institute's Nonprofit Data Project[3] began to tackle the problem of incomplete and hard-to-access nonprofit data. In 2011, it commissioned data experts, Beth Simone Noveck and Daniel L. Goroff, to study the Form 990. Their report, *Information for Impact: Liberating Nonprofit Sector Data* (2013),[4] articulated the benefits of open 990 data for research and scholarship, for nonprofits themselves, and for transparency and accountability purposes (e.g., giving charity regulators better tools to fight charity fraud and abuse). Two key recommendations were a universal electronic filing (e-filing) requirement for all IRS Form 990 editions, as well as a call to require the IRS to release the data to the public for free and in machine-readable format. The authors described how the information contained in the 990s could be more useful if it were public and open:

> *We argue that open 990 data may increase transparency for nonprofit organizations, making it easier for state and federal authorities to detect fraud, spur innovation in the nonprofit sector and, above all, help us to understand the potential value of the 990 data. … The sector deserves comprehensive and computable data that can be openly aggregated, searched, checked, and analyzed. (Noveck & Goroff, pp. 2–3)*

The recommendations from the 2013 report were actively promoted by the Aspen Institute's Nonprofit Data Project and its partners, including Guidestar, the Foundation Center, the Urban Institute's Center on Nonprofits and Philanthropy, the Lilly Family School of Philanthropy at Indiana University, and the Center for Civil Society Studies at Johns Hopkins University. The proposals resonated with the field and policymakers. By 2014, the report's primary recommendations appeared in President Obama's executive budget, a key Republican tax reform proposal, and a report from the U.S. Government Accountability Office. Notably, a successful federal lawsuit was filed by the open data advocacy group, Public.Resource.Org. Following this victory, news organizations such as *The Washington Post* and *The Chronicle of Philanthropy* quickly took advantage of the court's decision that forced the IRS to release Form 990 data via Freedom of Information Act requests. The IRS then engaged with members of the 990-user community to start testing data sharing formats. Finally, in June 2016,[5] the IRS took the historic step of releasing to Amazon Web Services — for free — machine-readable data of every field on the 990, 990-EZ, and 990-PF series of forms and schedules. At the

[2] Elizabeth Boris, Bill Levis, and Chuck McLean, to name a few of the leaders in these efforts.

[3] The Aspen Institute's Nonprofit Data Project was created in 2008 with funding from the C.S. Mott Foundation and the initial participation of Guidestar, the Foundation Center, the Urban Institute's Center on Nonprofits and Philanthropy, the Lilly Family School of Philanthropy at Indiana University, and the Center for Civil Society Studies at Johns Hopkins University. Working with members of the 990 data community, the project helped to spawn the Nonprofit Open Data Collective. These efforts help to strengthen the work of nonprofits on the ground, enhance transparency, and provide critical information and knowledge to those who work for and with nonprofits such as donors, government officials, members of the public, charity regulators, and scholars.

[4] https://www.aspeninstitute.org/publications/information-impact-liberating-nonprofit-sector-data/ A second edition was published in September, 2013, https://www.aspeninstitute.org/wp-content/uploads/files/content/docs/pubs/Information_for_Impact_Report_FINAL_REPORT_9-26-13.pdf

[5] IRS opens up Form 990 data, ushering nonprofit sector into the age of transparency: Sunlight Foundation

time of release, this data feed represented only organizations that filed their returns electronically, which was about 60% of all nonprofit filers of those three forms.

While the IRS release of e-filed data was crucial, much work remained. A community of 990 data experts from across the country met at the Aspen Institute to hold "datathon" sessions that began the labor-intensive process of cleaning and converting e-filed 990 data into more accessible public spreadsheets. Furthermore, paper-filed tax forms still remained unavailable as open data, a fix that could only be achieved through legislation. Policy efforts by the Aspen Institute and partners continued, eventually resulting in the passage of bipartisan legislation, the Taxpayer First Act of 2019, which mandated e-filing for all nonprofits required to file returns, as well as releasing 990 data to the public in a machine-readable format for free by the IRS.

E-filing requirements for most organizations[6] took effect in 2021. Form 990-EZ filers[7] were granted an additional year to transition to mandatory e-filing, and most will be doing so in 2022[8]. In late 2021, the IRS announced that it would "no longer update the Form 990 Series data on Amazon Web Services."[9] Instead, the IRS will make this data available solely on its own Tax Exempt Organization Search webpage.

Roughly ten years after the 2013 report, collective action has realized a key portion of the vision: the 990 data have been liberated.

What Open 990 has Meant to the Field

By nature, open data projects like Open 990 are massively collaborative endeavors — and it is from these collaborations that they get their strength and maximize their impact. As Markets for Good noted in a 2012 report titled, *Upgrading the Information Infrastructure for Social Change*:[10]

> *"Individually [data] is useful. Cross-referenced and connected, it is potentially transformative in its ability to allow stakeholders to make better decisions about budgets, strategies, services, policies, and more." (p. 7)*

The 990 data provides an exceptional amount of detail about organizations of all sizes and mission areas — a wealth of information far beyond what is available in corporation filings with state registration entities, for instance. This dataset is invaluable for journalists, regulators, researchers, and others who seek to understand the state of the 1.9 million nonprofits in the U.S. and the upwards of 12 million Americans who work for them. As shown in this section, **Open 990 users have, indeed, connected that information to other**

[6] https://www.irs.gov/e-file-providers/e-file-for-charities-and-non-profits

[7] The Form 990-EZ is used by nonprofit organizations with gross income of more than $50,000, but less than $200,000 — and with assets of less than $500,000.

[8] For more information on the electronic filing requirement, see the Aspen Institute brochure, "Mandatory 990 E-Filing: An Introduction." https://www.aspeninstitute.org/publications/opening-the-990/

[9] IRS. (2021, December 16). News release, para. 2. https://www.irs.gov/newsroom/irs-makes-tax-exempt-organization-search-primary-source-to-get-exempt-organization-data

[10] Markets for Good. (2012, Fall). *Upgrading the information infrastructure for social change.* https://digitalimpact.io/wp-content/uploads/2014/05/MarketsforGood_Information-Infrastructure_Fall-2012_.pdf

datasets for regulatory, investigative journalism, anti-fraud, public watchdog, community relief, nonprofit improvement, and research purposes.

The Open 990 dataset has given the field new insight into areas of nonprofit activity not available before:

- With access to every electronically-filing nonprofit and every field on the 990 form series, Open 990 provides a **detailed baseline of the depth and breadth of the nonprofit sector**. One researcher observed that instead of approximately 200 variables available for the selected nonprofits in the pre-2016 dataset, the Open 990 data series provides more than 4,000 variables across the multiple forms. This depth and breadth also encourages nonprofits to develop their own peer comparisons — "How does our cash on hand compare to other nonprofits of similar size," "Is our endowment larger or smaller than other foundations with similar missions," "How do our senior staff salaries compare to other nonprofits in our region?" — without relying on proprietary data sources from third parties.

- Open 990 data allow any user to **look at millions of nonprofit organization filings at once**, instead of accessing the static image files for a single organization at a time. In 2014, for example, a study detailing the economic impact of nonprofits in California required manual data entry on top of paying a $25,000 charge for access to digitized data for 40,000 public charities. With the Open 990 data, a 2019 repeat of the project avoided digitization charges and expanded to include all California electronically-filing nonprofits.[11] Similar work in Michigan expanded a study of private foundations — where data were collected manually from 48 Michigan foundations — into a study that included more than 67,000 private foundations across the nation.[12] Data providers and researchers are able to do things at a scale and speed that is unbelievable compared to just five years ago.

- Additional data **now include information from the forms that were inaccessible before**, especially on the schedules. Supplemental financial statements, summaries of donor advised funds (DAFs), detailed information about both nonprofit schools and hospitals, non-cash contributions, related party transactions, lobbying and political activity, and mergers and dissolutions are examples of the information now available with the Open 990 dataset. Further, these schedules are available at nearly a universal scale, instead of only selected items for selected nonprofits in a given sample. As one user noted, "I always tell people Open 990 nonprofit data is the best-kept secret in organizational studies."

- Machine-readable data means that Open 990 is a **key tool for ensuring good governance and transparency across the sector**. Charity regulators no longer spend weeks or months perusing individual PDF files, manually searching for patterns of suspicious organizational behavior. Open data means that questionable charity schemes — especially regarding fraudulent fundraising — are investigated and shut down faster than before. As one regulator said, "Relying on investigative journalists and whistleblowers is not as effective a monitoring device as having access to a universe of filings." Publishing machine-readable data also reduces errors that were previously introduced to the dataset via optical character recognition or manual data entry, processes previously used to transform image files into searchable information. Open 990 data users now see the data exactly as they were submitted by the nonprofit to the IRS, and the transparency encourages nonprofits to pay closer attention to what they are reporting and sharing with the public and with regulators.

[11] https://calnonprofits.org/publications/causes-count

[12] https://www.michiganfoundations.org/resources/payout-study

*Advances in computer-based text analysis and natural language processing, coupled with more complete information from the 990, opens **new avenues for research and analysis.***

- Advances in computer-based text analysis and natural language processing, coupled with more complete information from the 990, open **new avenues for research and analysis by enabling users to search open-text fields**, especially contained in Schedule O's "supplemental information" section. Many text fields, particularly from the forms' schedules, were not transcribed in the past, meaning this information is a new area for exploration. With new technologies, data processors can extract the mission statement provided by a nonprofit and update an organization's classification in ways that were impossible or impractical before. One Open 990 user noted that technological advances enable faster and more focused searches for nonprofits, such as, "Find all the nonprofits that focus on education of women or girls," or "Find all nonprofits that provide services to aging caregivers."
- **Value-added datasets and applications are now developed faster** because Open 990 is readily available at scale for technically-proficient data processors. Quite simply, data providers can more efficiently move to the next level of analysis because the baseline work is already done. Vanguard Charitable, for example, previously dedicated substantial time to uploading and parsing image files directly from the IRS, augmented by manual data entry by Vanguard Charitable's own staff as donor requests were received. Now Vanguard Charitable relies on a third party that directly accesses the Open 990 dataset, ensuring that electronic files are loaded into the DAF-provider's website shortly after release from the IRS. The increased speed of access and more universal coverage allowed Vanguard Charitable's internal staff to redeploy to value-added applications, including developing its Nonprofit Aid Visualizer (NAVi)[13] during the early days of COVID-19 pandemic response — a scale and speed that was unattainable prior to 2016.

To more explicitly demonstrate the examples above, the research team compiled examples of how 990 data has been deployed from the perspectives of various stakeholders. Readers are encouraged to visit https://johnsoncenter.org/resource/use-cases-from-publicly-available-990-data for additional examples.

Challenges

The sheer volume of data available in the Open 990 is both one of its greatest assets and a "baked in" challenge. For example: nonprofit organizational data is collected and reported in many ways; by multiple actors at both the state and federal level; and shared back with the public through various companies, datasets, and networks. In addition, each form type asks for different information, and there may even be inconsistencies within one form type from year to year.

While unlocking the data was essential, public access to the Open 990 files was only the start of the work necessary for true public use.

[13] https://www.vanguardcharitable.org/navi

Ease of use is a real concern.

The total size of the Open 990 data means it is generally impractical for individuals acting alone to download, manage, or analyze the information. To review an electronically-filed 990, an individual journalist, regulator, or member of the public would need to complete many file structure, data acquisition, and data extraction tasks to use the raw feed. Industrial scale storage and computing power — found at a university, large nonprofit, or government entity — are necessary to fully download and manipulate this data, especially if the goal is to compare multiple organizations or track changes over time.

As one regulator noted, in many states there are a handful of staff in an Attorney General's office that may work part-time on charity issues and part-time on consumer protection issues. These regulators are highly-skilled attorneys — not data acquisition specialists or database analysts. They need a data file that is already structured, easily searchable, and readily available on both a trend-level and individual organization basis. While detailed information about a single organization is readily available from sources such as Candid or ProPublica, it is much more challenging to get a dataset of information for hundreds or thousands of filers at a time. Previously, NCCS and other entities maintained "grab and go" core data files that helped serve this role.[14] In the Open 990 world, a user would go to one of the nonprofit academic research centers or consultants to access a similar file — or be prepared to download the Open 990 data, extract both the fields and nonprofits of interest, and construct a dataset before even starting the analysis. While the breadth and depth of Open 990 is revolutionary, it is a very "high effort" dataset for an individual user trying to have direct access.

In contrast, consider the difference between the raw, downloadable data from the Federal Elections Commission (FEC), compared to the easy-to-navigate "point and click" interface from the FEC itself.[15] Then compare the FEC's simple interface to the wealth of data and additional context from the website, OpenSecrets.[16] While election data is less extensive than the massive amount of data available in the 990 data series, and election data has been publicly accessible for more years, the FEC's example shows the possibility of a similar federally-collected set of data that has been cleaned and curated for the general public.

In addition, Open 990 lacks a comprehensive data dictionary. The IRS publishes the Open 990 data in a raw form, without an updated list of the field names, field types, and expected or valid values in the fields. The lack of a "schema," which provides a blueprint for how the underlying data are organized, makes it challenging for direct users to determine whether the same information is in the same field from year to year, or to determine which portions of the form are contained within each component of the raw data file.

Perennial gaps and challenges remain.

The complexity and scale of the nonprofit sector presents a challenge to data collection and timely distribution of information. The 990 data is incredibly detailed, providing a wealth of insight into each individual organization. However, that also means we have not yet found sufficient ways to systematically gather and share all the individual elements that could ultimately enable the sector to do better work and generate more equitable and inclusive outcomes.

[14] See, for example, https://nccs-data.urban.org/data.php?ds=core

[15] https://www.fec.gov/data/

[16] https://www.opensecrets.org/

Remaining gaps include:

- **The timeliness of data.** The IRS does not release the data from electronically-filed 990s on a predictable or particularly timely schedule. Up to nine to 12 months can pass between the time an organization files a 990 and the first appearance of the data in the open data repository. This makes it difficult for foundations and nonprofits seeking to respond promptly to changing needs and emergencies (such as a global pandemic or other natural disaster, for instance) to respond quickly and reach all of the organizations who might need, or be able to provide, assistance.
- **Incomplete information.** Even though electronic filing is mandated, some nonprofits provide required information in alternative formats that make electronic analysis more difficult. A common example is a foundation that provides its grantee list in an unstructured text attachment, instead of in the structured fields found on Schedule I, or a nonprofit that does not cleanly list its directors and compensation information in Schedule J. Missing data is also an issue, when fields are left partially or completely blank.
- **Mismatched input fields across the various 990 forms and potentially related datasets.** One advantage of open data is the ability to combine and/or compare large scale datasets from one source with others. However, inconsistencies in field naming, field definitions, and collected information requires a significant lift from human analysts to make those comparisons possible. The four primary 990 forms (990, 990-EZ, 990-N, and 990-PF) do not make the same input fields available to each category of filer[17] — and even when the same fields are requested (such as revenue or expenses) they appear in different parts, sections, and line numbers of each form. For example, while the number of staff and volunteers is requested on the 990, similar information is not requested on the 990-PF. Similarly, while a grantee's Employer Identification Number (EIN) is included as part of the detailed listing of a community foundation's grants on the 990, there is no EIN field in the detailed listing of a private foundation's grants.
- **Not all nonprofits operate a single organization from a single location.** Existing data ascribes all organizational information to a single location on a map — a "home base," of sorts, for each nonprofit. However, many organizations operate from more than one location (e.g., a nonprofit daycare with many locations in a region), operate more than one branch (e.g., a federated network like the United Way), or share an EIN with other entities or programs that have their own public identity (e.g., the Tides Foundation provides fiscal sponsorship for Emerging Practitioners in Philanthropy and others).
- **For many of the sector's big questions, the data or linkage is not available.** While 990s provide an overall view of an organization at a moment in time, there is a significant amount of nonprofit sector-related data that we do not have access to. Examples include:
    - a more robust and more frequent release of nonprofit employment data from the Bureau of Labor Statistics, as well as more regular updates to the Volunteer Supplement of the Census Bureau's Current Population Survey;

[17] As one observer noted, this is an example of "no good deed goes unpunished." While it is logical for the IRS to require less information from a 990-EZ filer, the lack of information for these smaller nonprofits can make the organization seem "less real, less stable, and/or less transparent" compared to a full Form 990 filer. This informational difference is acute from a donor perspective, for example, in deciding between two nonprofits.

- easy "round trip" information — e.g., the pathway necessary to find a nonprofit, find who funded that nonprofit, then find information about those funders and other grants from those funders to similar organizations;[18]
- up-to-date assignment of subsector codes (such as the National Taxonomy of Exempt Entities) for every nonprofit, which can then be reliably mapped to other categorizations such as the North American Industry Classification System (especially for nonprofit subsectors where there are ready comparisons to the for-profit sector, such as health care entities); and,
- direct links from federal-level data (e.g., IRS data) to state- and local-level data, such as charity and/or fundraising registrations and corporate organization, service territories, or community-wide regulatory and informational filings (e.g., a hospital's latest Community Health Assessment or Community Health Improvement Plan).

These are not merely academic concerns. For example, while the EIN was required as part of the application for a Paycheck Protection Program (PPP) loan, the EIN was not included as part of the data released by the Small Business Administration in the early days of the COVID-19 pandemic response. This omission effectively blinded regulators, journalists, and nonprofit leaders alike in determining whether PPP loans were effectively distributed across the nonprofit sector, let alone nonprofits in one state compared to another.

Inspiration and lessons learned from other Big Data projects

As noted above, the Open 990 project is one of many open data projects currently in operation. Over time, these projects have yielded inspiring stories and cautionary tales to help guide the work of Open 990 efforts going forward.

- **The real power comes when datasets are combined with one another — and with human intelligence.** Individual datasets have great value as tools for understanding aspects of a certain topic. When datasets are combined, however, with one another and with human meaning-making, their potential for creating actionable value increases exponentially. For example, the usefulness of 990 data improves when it is matched to:
  - state-level entity registrations — for example, linking the federal EIN to the unique state-level identifier where the entity is registered/incorporated;
  - American Community Survey and census data, especially related to the population within a community-focused nonprofit's service area; and
  - datasets that show participation in federal activities, such as contracts and grants from Federal agencies to nonprofit organizations as well as nonprofits that received Paycheck Protection Program assistance during the COVID-19 pandemic.

  However, as we move from "open data" to "big data" — that is, aligning and combining different datasets in order to reveal larger insights beyond the basic inputs — we run the risk of producing misleading results if we remove human judgment from the equation. Big data is not a substitute for

[18] Or, similarly, find a nonprofit, find the directors of that nonprofit, and then see what other nonprofits that person is associated with either as a director, contractor/vendor, or employee.

human experience and judgment — it should augment and inform professional judgment. The creators, administrators, and users of open data projects must take implicit bias and discrimination into account and actively work to mitigate the risk of unintended harms.

An example of this is visible in the case of NarxCare, a big data medical software program intended to identify individuals at risk for opioid abuse (Szalavitz, 2021). NarxCare combines data from a person's medical and prescription history, public emergency services data, and criminal justice data, among other sources, to compute an individual's so-called Overdose Risk Score. However, because the system can retrieve data from multiple sources linked by a person's name and date of birth, related items — such as veterinary prescriptions — have resulted in a skewed score and denial of treatment for patients in need of real help.

Similarly, problems may arise as natural language processing and artificial intelligence are used to parse the free text fields. Older and/or larger nonprofits may write more nuanced and richer descriptions of their missions and programs, while smaller and/or younger nonprofits may use more straightforward and/or shorter descriptions. As machine-learning techniques analyze the text fields, will this do more to lift up previously underfunded nonprofits — or simply reinforce existing structures and size biases of more well-funded organizations?

Big data is not a substitute for human experience and judgment — it should augment and inform professional judgment. The creators, administrators, and users of open data projects must take implicit bias and discrimination into account and actively work to mitigate the risk of unintended harms.

- **The more a dataset gets used, the more flaws are identified — and the result can be higher quality data.** Joining datasets is hard. The more frequently data is used, combined, and analyzed, the more flaws in a dataset come to light. APIs break, scripts break, entity names vary across datasets (e.g., differences in capitalization, punctuation, abbreviations). Corporate names can change, or an organization can gain or shed new locations and branches without updating their EIN. Data dictionaries are not always provided with the data, and when they are provided, frequently they are out of sync with the most recent dataset. Different states collect and make available different types of data from different years and in different formats.

  This is not a new issue, nor a challenge inherent only to open data projects (see also Fisher et al., 2002; Gordon et al., 2007). As these projects put data and analytical capacity into the hands of more and more people, the potential for incorrect, mismatched, or otherwise flawed data to cause harm grows — especially when regulators, investigative journalists, and others are looking to draw conclusions from the data available. Without a common system for error-checking and data cleaning, this pitfall will continue to cause trip ups.

  But while open data can be like having too many cooks in the kitchen, the good news is that every cook brings their unique recipe to the table. With every additional pair of eyes on a dataset, the more people and organizations find ways to clean and improve the data and put it to greater use. The open nature of the project means that entities are effectively crowdsourcing the work of improving the data's quality over time.

- **Ongoing maintenance and updates are as important as the initial release — but are not as interesting to funders.** Once the initial work is completed and published, an ongoing series of tasks — data acquisition, data cleaning, historical archives, data evangelizing — can keep whole teams busy for years. Open data is merely freely available; its usefulness is the base of a mountain that can soar to unimaginable heights — but only if effectively resourced. But "maintenance" is not the easiest sell.

  Our discussion with the organization OpenSecrets, for example, related the story of congressional financial disclosures. For several years, OpenSecrets had posted the image files of personal financial disclosures to its website. However, as the Enron company began to fail, it was very difficult for journalists to identify which members owned direct stock in the company. In the worlds of campaign finance as well as journalism, when something happens observers need to have the data on hand, not be required to generate the data in the middle of an election cycle or cataclysmic financial situation. The "grunt" work of ensuring that as many fields as possible are as searchable as possible, every time a new data file is released, takes ongoing maintenance and operational support from funders, subscribers, or donors.

Recommendations

The availability of Open 990 data has led to a rapid increase in the knowledge of the sector because observers are now able to analyze data about the sector at scale, nationwide, and over multiple years. As research continues, linking Open 990 data to other datasets, as well as deploying new machine-learning techniques and artificial intelligence as part of data analysis, will yield even more knowledge. But this massive dataset has also highlighted challenges, some inherent in the data, while other challenges emerged from the way the data is currently made available.

These challenges are not insurmountable — in fact, they point towards potential solutions that encourage collaboration and shared meaning-making across the nonprofit sector and alongside partners in other sectors and communities. To move the Open 990 Project forward, we have to be our own greatest advocates.

These challenges are not insurmountable — in fact, they point towards potential solutions that encourage collaboration and shared meaning-making across the nonprofit sector and alongside partners in other sectors and communities.

Collaboration is essential to create a coordinated Open 990 data ecosystem.

Today, Open 990 data is processed via an informal network of volunteers, researchers, and both commercial and nonprofit data providers. The need for industrial scale computing, for instance, opens the door for consumers to use third-party intermediaries such as Candid's GuideStar or ProPublica's Nonprofit Explorer to parse and present information about single nonprofits — or use fee-for-service offerings from entities such as Candid, DataLake, or Citizen Audit to create custom data extracts of multiple organizations across years. These intermediaries provide small-scale peers and those without in-house data expertise with the data and insights they need to target programs, evaluate impact, and identify collaborators. Similarly, some

researchers and data experts have published code to GitHub.com which gives anyone easy access to the raw IRS 990 data at no cost, and provides helpful step-by-step instructions.

But what is missing is a coordinated tier of supports with clear access points. For example, a coordinated ecosystem would make available:

- A base level offering with access to the raw data feed from the IRS, similar to what exists now, but adding a regularly updated code book and stable data release schedule.
- A level 1 offering that de-duplicates and cleans the raw data feed, offering the resulting complete dataset in multiple formats such as SQL, JSON, and XML for ease of use.
- A level 2 offering that translates the level 1 dataset into tables containing selected variables, pre-sorted by state and/or subsector, provided to users in SQL and XLSX formats for widest accessibility ("grab and go" datasets of the most commonly used variables).
- A level 3 offering that summarizes the existing data feed on a monthly, quarterly, and annual basis for trend analysis in XLSX format for broad availability.

This ecosystem creates multiple places for engagement from commercial and nonprofit data providers alike, while also increasing the accessibility and utility of the data files for custom "mash ups" of data with other, related datasets. By reducing or eliminating the time necessary to obtain and clean the raw Open 990 data, a coordinated approach can save the sector an enormous amount of time and resources, avoiding a massive duplication of efforts while also increasing the speed of innovation.

The IRS can take essential steps to improve the utility, consistency, and accessibility of the existing Open 990 data.

Each of the items below has been previously noted by the Aspen Institute's Nonprofit Data Project through public comment to the IRS. While each of these steps appear relatively small in isolation, collectively these administrative changes are essential to improving the public's access to and use of Open 990 data.

- Provide a regular schedule of data release dates to improve analysis and use, including the expected lag time between the date of filing and the date of public availability.
- Publish the data schema for the existing raw data feed and create/improve data indices.
- Engage the research community for suggestions about improving machine-readable data accessibility and categorization in text fields, such as in Schedule O and elsewhere.

Provide support for the IRS tax-exempt staff — especially the data processing team.

Because the 990 forms generally do not generate revenue for the U.S. government, there is less impetus for the IRS to dedicate resources to ensuring that the data they receive and release to the public is accessible and useful. The IRS staff dedicated to this work has been historically underfunded, especially relative to the size of the nonprofit sector and its share of both employment and wages in the United States.

Require federal agencies to disaggregate nonprofit data.

The single best action to quickly advance knowledge throughout the nonprofit sector would be to require every federal agency to disaggregate its data for the nonprofit sector whenever it publishes data at a sectoral or industry level. This includes requiring the U.S. Department of Labor to release statewide and sub-sector summaries of nonprofit employment and wages on a quarterly basis.

Encourage or require existing Open 990 data providers to be more transparent.

The informational nature of the IRS Form 990 also means that there are few consequences for a nonprofit if it provides incorrect data or omits data in its filings with the IRS. There is no standardized — and publicly available — system for error detection and correction of the data contained in the Open 990 data feed. This means that any entity wishing to use the Open 990 datasets is left to construct their own systems for error checking and correction — or to simply take the raw Open 990 data feed "as is" and accept errors and inconsistencies as a "known risk." Those who use 990 data and develop solutions to problems should be encouraged to share their knowledge with others. In addition, the IRS should be required to publish the error detection and correction business rules it uses in processing the Form 990 returns. Coupled with this, creating incentives and outlets for major Open 990 data providers to publish their error detection and correction business rules would allow users to examine additional checks and balances they could utilize depending on their research needs.

Conclusion

When we put data into the hands of individuals and organizations working for the public good — and when we come alongside them with the tools and expertise necessary to easily access, use, and track that data over time — we create the conditions that encourage a sea change in the social sector. We build a foundation of transparency and clarity that can encourage public trust in nonprofits (Edelman Data & Intelligence, 2021) and empower the kinds of community-grounded organizations and programs that more efficiently solve problems and advance an equitable future.

The availability of Open 990 data allows new insights into the health of the nonprofit sector as well as the pathways along which funds, resources, and people travel across organizations and communities. Five years after the release of Open 990 data, this initiative is encouraging enhanced collaboration, trust, and accountability through active use of the dataset. With additional coordination — matching the 990 data to other existing public datasets, and clarification from the IRS and other stakeholders on the underlying data structure and cleaning procedures — Open 990 data has begun to generate important insights and support a thriving civil society.

Appendix: Tools for Accessing the Open 990 Data

Examples are presented in each category **in alphabetical order by host institution/program**.

Where to access the Open 990 raw data

- Prior to January 1, 2022, the Internal Revenue Service Registry of Open 990 Data on Amazon Web Services had been the official source for data released by the IRS. In 2016, the IRS began releasing e-filed 990 data, approximately 60%–65% of filers at the time, to Amazon Web Services (AWS) for public use, free of charge. It is available for years 2013 to 2021. The data has been posted as XML files.[19]
- Starting on January 1, 2022, the Internal Revenue Service Tax Exempt Organization Search Bulk Data Downloads page is now the main repository for 990 downloads in both PDF and XML formats, organized by year and month, depending on the format.

How to access and use Open 990 raw data

- The Registry of Open 990 Data on Amazon Web Services has a list of tools and applications for using the raw 990 data.
- The Nonprofit Open Data Collective on GitHub.com serves as a shared repository of tools, tips, and resources for cleaning and processing 990 data, including a helpful link to the ARNOVA 2017 workshop, Open Nonprofit Data, and a place where scholars and others can share research. The website also includes a long list of other helpful websites and datasets for nonprofit research, such as a tool that can tell whether a nonprofit has a recent e-filed return and a way to classify nonprofit mission statements, and a way to convert the list of board member names into a more user-friendly format.
- IRSx.info includes information that helps 990 data users convert raw XML files into named variables, with each data point mapped to its location on the IRS Form 990 series filings.

Examples of sites that use and leverage Open 990 data

- American University — The Accountability Project gives researchers and journalists tools for searching across public datasets, including 990s and more than 1.4 billion public records. In addition to 990 data, the site includes data on business ownership, public employees, medical facilities, voter registration, and government spending.
- Candid provides research on trends in the nonprofit sector and gives users the ability to look up individual nonprofit and foundation tax forms. Candid's GuideStar provides free access to individual nonprofit and foundation tax forms, and FDO Quick Start contains free profiles of grantmakers. Both platforms offer more in-depth and customized information upon subscription as well as reports on nonprofit compensation. Nonprofits may earn seals of transparency on their online profiles to help attract donors.

[19] As of December 31, 2021, this site has been replaced by the IRS bulk data download site in the second bullet. Starting January 1, 2022, the IRS site will contain historical archives as well as receive updated data, while the AWS site only maintains historical archives.

- Cause IQ provides tools to help a variety of businesses, nonprofits, and others grow their nonprofit client base, research organizations, and find funding. Users can search nonprofit organizations based on location (city or state) and category (colleges, nursing homes, animal organizations, etc.). This platform also contains 990-based research reports, such as a guide to COVID-19 Relief Funds.
- Charity Navigator assists donors by providing them with free access to data, tools, and resources that guide philanthropic decision-making, including advanced searches of nonprofits by issue area, rating, and more. Charity Navigator's Encompass Rating System uses open 990 data to rate a growing number of nonprofits across several key indicators, including: Finance & Accountability, Impact & Results, Culture & Community, and Leadership & Adaptability.
- CitizenAudit.org has a searchable database of Form 990s, including access to 15 years of Form 990 disclosures. Users can easily search for keywords inside the IRS filings, and more custom needs can be met, from alerts and notifications for the latest filings to the creation of tailored lists. CitizenAudit charges an annual subscription fee though basic searches are free.
- DataLake Nonprofit Research provides custom national, statewide, and regional reports and datasets for a fee. They offer access to 30 years of Form 990 data and National Taxonomy of Exempt Entities (NTEE) classification of most tax-exempt 501(c) organizations.
- Foundation Mark provides indices and reports on the performance of foundation investments.
- GivingTuesday Data Commons provides access to e-filed 990 data and a host of resources for analysis of 990 and other data, with free registration. The site includes data visualizations and dashboards, as well as links to code, data cleaning files, scripts, and more.
- Grantmakers.io provides foundation profiles and a tool for searching grants.
- HospitalFinances.org is a platform by the Association of Health Care Journalists that makes nonprofit hospital finances easier to access, search, and analyze. Searches can be conducted by entering the name of facilities or individuals and results will show information on finances, charity and indigent care, compensation, and more.
- The open990.org website contains nonprofit profiles as well as resources from other researchers and organizations. The site also contains free datasets on topics such as hospitals (Schedule H data) and executive compensation.
- ProPublica's Nonprofit Explorer enables searches of nonprofit tax returns through a variety of filters, including state, major nonprofit category, organizational type, people, and full-text search. This site also includes PDFs of Federal Single Audits for nonprofit organizations that spent $750,000 or more in federal grant money in a single fiscal year.
- NAVi – Nonprofit Aid Visualizer and NAViHH - Nonprofit Aid Visualizer for Hunger and Homelessness from Vanguard Charitable provides search tools using filters focused on finding nonprofits to donate to support COVID-19 pandemic relief, hunger, and homelessness. The sites include nonprofit profiles and maps of the distribution of nonprofits.

Other sources of information about the nonprofit sector

Beyond the Open 990 data, there are a number of other data sources and resources that share information about the nonprofit sector. Two of the most important are:

- The Internal Revenue Service's Charitable and Exempt Organization Statistics
    - The Annual Extract of Tax-Exempt Financial Data, which includes selected financial data from all Forms 990, 990-EZ and 990-PF filed in a calendar year
    - The Business Master File, updated on a regular basis, which lists all active nonprofits by state and region
    - Exempt Organizations Form 1023-EZ Approvals, which provides an annual summary of all nonprofits that were created with the streamlined application for 501(c)3 status
- National Center for Charitable Statistics at Urban Institute
    - NCCS Core Files, which contain selected information from all organizations filing a Form 990 or 990-EZ. NCCS also maintains an excellent archive of historical core files going back to 1989.
    - The historical archives of Business Master Files and Statistics of Income datasets

Other key nonprofit data resources include:

- Arizona State University's Lodestar Center for Philanthropy and Nonprofit Innovation
- Aspen Institute's Program on Philanthropy and Social Innovation, including the 990 e-filing guide
- Association of Fundraising Professionals — The Fundraising Effectiveness Project
- The Generosity Commission published an appendix of U.S. databases about philanthropy (June 2019)
- Grand Valley State University's Dorothy A. Johnson Center for Philanthropy
- Independent Sector regularly publishes reports on the health of the U.S. nonprofit sector
- Indiana University's Lilly Family School of Philanthropy
- Johns Hopkins University — The Center for Civil Society Studies
- National Council of Nonprofits publishes reports on the economic impact of the nonprofit sector
- Stanford Center on Philanthropy and Civil Society
- The University of Maryland's School of Public Policy (Philanthropy & Nonprofit focus)
- The University of San Diego's Nonprofit Institute
- The University of Texas at Austin's RGK Center for Philanthropy and Community Service
- The Urban Institute's Center on Nonprofits & Philanthropy