1996

State Nonprofit Data Bases: Lessons from the California Experience

Richard J. Orend
University of San Francisco

Michael O'Neill
University of San Francisco, oneill@usfca.edu

Connie S. Mitchell
University of San Francisco

Recommended Citation

Orend, R. J., O'Neill, M. and Mitchell, C. S. (1996), State nonprofit data bases: Lessons from the California experience. Nonprofit Management and Leadership, 7: 447–454. http://dx.doi.org/10.1002/nml.4130070410

RESEARCH REPORTS

State Nonprofit Data Bases: Lessons from the California Experience

Richard J. Orend, Michael O'Neill, Connie S. Mitchell

IN JULY 1990, the Institute for Nonprofit Organization Management at the University of San Francisco launched the California Nonprofit Database (CND) Project. One purpose of CND was to present a comprehensive, in-depth picture of California's nonprofit sector. Additionally, CND was intended to provide scholars and policy analysts, funders, individual nonprofits and umbrella organizations, state and local government agencies, the media, and the public with information on various aspects of nonprofit activity in the state. A general purpose was the promotion of increased communication and collaboration among nonprofit data base efforts at the state and national levels.

CND published *California Nonprofit Organizations 1995,* a two-hundred-page statistical report, in October 1995. The report is available from the Publications Department, Institute for Nonprofit Organization Management, USF, 4306 Geary Blvd., Suite 201, San Francisco, CA 94118; (415) 750-5183; <KOZIOL@USFCA.EDU>. Many smaller and more specialized reports and analyses have been produced, and others are currently under way.

This article briefly summarizes some of the major lessons learned during six years of working on the CND project. We hope these lessons will be of value to others who are developing or thinking of developing state or regional nonprofit data bases, as well as other scholars and practitioners interested in the quantity and quality of data on nonprofit organizations. The lessons are organized under the following questions: (1) What should be included in the data base? (2) What are the data sources? (3) How is data quality ensured? (4) How should the data base be structured? (5) How is the information used, and by whom? (6) What will it cost to develop and maintain the data base?

What Should Be Included in the Data Base?

This question has three parts: What organizations should be included? How should these organizations be categorized? What data should be included? The answers to these questions are determined both by the objectives of the developers and by data availability.

What Organizations Should Be Included?

The Internal Revenue Code (IRC), under Section 501(c), lists twenty-five different types of organizations as exempt from federal corporate income tax. The majority of these fall under the 501(c)(3) code: religious, educational, charitable, and other such organizations. Other IRC categories include social welfare agencies, fraternal societies, business leagues, social and recreation clubs, and a variety of other groups. CND includes all types of organizations listed as exempt and classified under California state law as public benefit, mutual benefit, and religious nonprofit corporations. While some nonprofit data bases omit or deemphasize religious and mutual benefit nonprofits, we felt that their inclusion was essential to present a comprehensive and accurate picture of California's nonprofit sector.

CND includes all types of organizations listed as exempt and classified under California state law as public benefit, mutual benefit, and religious nonprofit corporations

CND includes about 120,000 organizations, nearly 75 percent of which are quite small, that is, they have no employees, annual revenues of less than $25,000, or both. Unincorporated associations and other voluntary groups are not currently included in CND due to the general unavailability of data; however, many have argued that such groups play an important role in society (for example, D. Smith, 1991, 1993) and constitute significant percentages in some nonprofit subsectors (for example, Kaple, Morris, Rivkin-Fish, and DiMaggio, 1996). Although only nonprofit organizations are formally included in CND, our 1995 report presented comparative data on for-profit and government agencies in several industries (for example, health care, education, arts and culture, and social services) in which the work of nonprofits must be seen within a larger economic context.

How Should These Organizations Be Categorized?

Currently, there is no universally accepted system for classifying nonprofit organizations. There are several systems in use or under development at the national and international levels, and many more in use at the state level. In developing CND, we have encountered the Internal Revenue Service (IRS) activity and purpose codes; the federal Standard Industrial Classification (SIC) system and its recent reincarnation, the North American Industrial Classification System; the National Taxonomy of Exempt Entities (NTEE), developed by the National Center for Charitable Statistics formerly at INDEPENDENT SECTOR and now at the Urban Institute; the California Registry of Charitable Trusts (RCT) coding system; the California Secretary of State (SoS) codes; information and referral services human services organization codes (for example, Sales, 1991); and nongovernment systems such as those used by the American Hospital Association and the National Catholic Educational Association.

CND dealt with the classification problem by using the best available data and, whenever possible, multiple codes. An example of the former is the fact that, for many purposes, the best data on California nonprofits comes from the quinquennial Census of Services Industries (CSI), conducted every year ending in a 2 or a 7 by the U.S. Bureau of the Census. CSI, like most federal studies including nonprofits, uses the SIC system. For example, California's RCT codes each nonprofit in its system by IRS activity codes, RCT codes, and California SoS codes.

A serious problem in most if not all of these systems is the inaccuracy and inconsistency with which some entities are classified. Many studies of IRS and NTEE classifications have found a high error rate (for example, Grønbjerg, 1994); our experience certainly confirmed this general finding. In sum, there are several important nonprofit classification systems in use, they are not interconvertible (crosswalks mitigate but do not solve this problem; see, for example, B. Smith, 1992), and the major systems do not yield a high degree of accuracy and consistency in the actual classification of nonprofits.

The classification problem is compounded by another characteristic of current systems: the use of only one or two codes to classify multifunctional organizations. This undoubtedly contributes to the misclassification phenomenon. Also, failure to identify all or most of the functions of multifunctional organizations means that data base users will miss potentially important information if they use one or two classifications as their only identification tool.

What Data Should Be Included?

Ideally, a nonprofit data base might include six types of information: financial, programmatic, administrative, personnel, historical, and demographic. Due to IRS interests reflected on Form 990, financial data are the most detailed and readily available. By law, this is public information. In California and some other states, 990 data are available in automated form. Programmatic and administrative information is generally not available in automated form but is invaluable for understanding what the organization does and how it is structured and managed. Personnel data are often difficult to obtain because of privacy restrictions. Historical data trace the foundation and development of the organization and are particularly useful in tracking change and the life-cycle patterns of nonprofit organizations. Demographic data may include organization location, information about facilities and property holdings, or other general descriptive information.

For the development of CND, the issue was and remains less a matter of defining what is desirable than of determining what is available. The strategy adopted was less a selection of interesting variables and more the development of a plan for collecting, over time, much of what is available. There are three levels of availability: automated data, data aggregated at a central point but not automated, and data residing at the individual organization. Due to time and cost factors, CND focused primarily on the first level. Also, the 1995 report used many secondary data sources.

What Are the Data Sources?

The foregoing discussion has addressed some of the issues involved in finding the appropriate data. Generally, most aggregated information is available from government agencies. The IRS provides a very small amount of financial information (assets and expenditures) from Form 990 tax returns. Many states require the submission of 990 and state forms to state agencies, and these forms are available in their entirety, although not always or completely automated. In California, the RCT provides a datatape containing about 80 percent of the 990 information and also provides access to individual returns for the remaining information (mostly narrative program information and officer names).

Generally, most aggregated information is available from government agencies

Other sources are widely dispersed and provide smaller amounts of data. The California SoS provides automated historical information (date of founding and so on) and the most current official addresses for all incorporated entities. SoS also provides hard-copy versions of the articles of incorporation for these organizations. The California Employment Development Division provides aggregated data on personnel, but this agency is legally restricted from releasing the data on individual organizations. Other states have less stringent limitations. The California Department of Education and the Postsecondary Education Commission provide data on private educational institutions. The Office of Statewide Health Planning and Development provides data on nonprofit hospitals and other health agencies. We can find out about aggregate property holdings of nonprofit organizations from the State Board of Equalization, but to get data on individual agencies we must contact assessment offices in fifty-three counties.

Private organizations are important sources of some kinds of data. State and national associations have information on their member agencies. National organizations concerned with nonprofit organization research, like the National Center for Charitable Statistics, are also building data bases and will have some state-level data available.

In developing a plan for data collection, several questions should be addressed: Where are the data located? Are the data public? Are the data automated? What will they cost? How accurate are they? How comprehensive are the data? It is a given that multiple data sources will be required. Because data accuracy cannot be assumed, multiple sources offer an opportunity for checking data reliability. In addition to the need to develop multiple data collection protocols to accommodate each different source, the disadvantage of multiple sources is the problem of linking different data sets.

Highly desirable data are often unavailable because of laws, regulations, or proprietary issues. For example, California's Franchise Tax Board has a full range of financial data comparable to the IRS Form 90 information, but laws restricting the public use of the state data are more stringent than the federal government regulations. Private associations also may restrict information about their members.

Ultimately, however, the greatest data restriction may be automation. Much of the information about program activities, organizational objectives, and names of officials and board members that is publicly available can only be obtained in hard-copy form, one organization at a time. The restriction is not availability per se but the cost of collecting and automating the data. For example, the IRS plans to produce computer images of all 990 tax forms but not digitize the information. These images, to be made available on CD-ROM, will still require key entry.

An example of a known systematic bias is the absence of financial data for religious organizations, which are not required to file tax returns

How Is Data Quality Ensured?

Data quality includes issues of reliability, validity, completeness, and comprehensiveness. In the development of CND, problems in each of these areas were encountered and addressed, although not all successfully. Reliability questions arise in the reporting and processing of data. The ambiguity of the 990 form, for example, can lead to highly unreliable reporting of certain types of revenue and expenditure information by nonprofit organizations (Froelich, 1996; Froelich and Knoepfle, 1996). The mere transfer of data from forms to computers can also create significant errors. The former problem can probably only be addressed by changing the 990. The latter problem is addressed by complete verification of data entry and the application of audit and logic check routines to the data base.

Validity is another issue that arises when different interpretations of the requirements are possible, as in the 990. It is often unclear just what the various responses to income and expenditure questions really mean. Also, statements about objectives and activities in public data bases are often conflicting and out of date, making it difficult to determine what organizations are doing, how to classify them, and how activities relate to financial and other information.

The 990 also can be used to illustrate the problems of incomplete data. Often the forms are not completely filled out or organizations required to file a 990 do not do so. To analyze data sets with many incomplete records, it may be necessary to impute values or, minimally, to recognize that incomplete data may represent a systematic bias in the findings. An example of a known systematic bias is the absence of financial data for religious organizations, which are not required to file tax returns. Potential unknown biases stem from the failure to complete all parts or lines of the 990.

Comprehensiveness refers to the number of organizations in the data base. One of the primary functions of a data base on organizations is to identify all or most of the organizations that comprise each category. In the current CND, we have identified California organizations that file 990s and those registered with the state and IRS, but we are only beginning the difficult task of identifying unincorporated associations. It is not possible for a data base developer to "correct" many of these errors, but it is the responsibility of the developer to be aware of the possible data problems and to make users aware of their magnitude and nature.

How Should the Data Base Be Structured?

This technical question addresses the interrelated issues of efficient data storage and maintenance, the software used to manage the data, user interface, and data file linkages. The most efficient way to store and manipulate the various files is to use a relational data base. This makes the process of adding new data and updating files easier. The critical element in this approach is linkage. Linkage is the ability to identify the same organization in two or more files because many users request data from multiple files and require presentation in a flat file for easy analysis or other use.

CND is based on data from a number of sources and each source is represented by a different file. The most widely used identification number is the IRS Employer Identification Number; however, many data sets do not use this number. To link organizations in the SoS file to the RCT file, for example, the California corporate number must be used. When that number is missing from the RCT record, a much more complex name-matching procedure must be used. Some files have no numbers and difficult, time-consuming (even when automated) name-matching procedures must be used to link records. Failure to develop and maintain highly reliable linkage variables seriously jeopardizes the quality and usefulness of the data.

How Is the Information Used, and by Whom?

Different users access CND data in different ways. Researchers generally seek larger files so they can conduct their own analyses. Foundations and other funders usually seek individual records or lists of specific types of organizations for review or comparison purposes. Service-providing nonprofits seek specific types of organizations for comparison (benchmarking), coalition building, or other purposes. Commercial users are most interested in particular categories of organizations or particular geographical areas and almost always seek current names and addresses.

Currently, CND is accessed only by Institute for Nonprofit Organization Management data-processing personnel. Although all of the data are available, we have found that it is best to create a specific file for each use. Requests from users are executed by the data base managers and finished files or hard-copy lists are given to users. These files have ranged from a few names and addresses to large data sets including a significant number of variables across several subsets of organizations. If there is heavy usage, this may become a burden for the data base managers, but it is probably easier than establishing open access via a modem-driven network. Eventually, a substantial portion of the data may be put on the Internet, but it is unlikely that Internet users will be permitted to download an entire data base. This last point relates to the question of cost for data use.

What Will It Cost to Develop and Maintain the Data Base?

CND has been supported by more than $500,000 in grants from the William and Flora Hewlett Foundation, W. K. Kellogg Foundation, Wallace Alexander Gerbode Foundation, William Randolph Hearst Foundation, James Irvine Foundation, Walter and Elise Haas Fund, Lilly Endowment, Chevron USA, and USL Capital. To date, a relatively small amount of funding has been generated from user fees.

Costs for developing and operating the data base go primarily to personnel for designing and maintaining the data base, data entry, responding to user requests, and marketing. Other marketing costs, such as the purchase of mailing lists and the cost of printing and mailing brochures, are a relatively small part of the total operating cost.

Hardware and software costs are relatively minimal but are dependent on the size and sophistication of the data base

Most data, purchased from state sources, are also relatively inexpensive, ranging from $110 per year for the RCT data (IRS 990 and a state form) to $2,000 per year for the SoS data. These costs will expand as sources expand and the costs of some new sources, such as the IRS 990 image data expected in late 1997, are revealed. The entry of unautomated data and the collection of new data from original sources will be very expensive.

Hardware and software costs are relatively minimal but are dependent on the size and sophistication of the data base. Currently, CND is housed on a university computer, so hardware costs include only the PC used for connecting and minimal software. A smaller data base, applicable to most states, could easily be housed and efficiently manipulated on a Pentium with a large hard drive. Software for accomplishing tasks like file matching, however, is more expensive—$3,000 for the PC version of the best system of which we are aware.

For keeping costs low and convenience, a PC system with something like dBASE or Access is clearly the best approach. As with CND, beginning with readily available automated data from national and state sources will provide a basic, useful data base at minimal cost. Personnel costs will drive the development, and even relatively small organizations can expect to devote a full-time data base manager for development, maintenance, and customer service. Special data collection efforts will require additional personnel.

RICHARD J. OREND is director of research for the Institute for Nonprofit Organization Management at the University of San Francisco.

MICHAEL O'NEILL is professor and director of the Institute for Nonprofit Organization Management at the University of San Francisco.

CONNIE S. MITCHELL is data base manager for the California Nonprofit Database Project at the Institute for Nonprofit Organization Management at the University of San Francisco.

References

Froelich, K. A. "The IRS 990 Return: Beyond the Internal Revenue Service." Unpublished manuscript, College of Business Administration, North Dakota State University, 1996.

Froelich, K. A., and Knoepfle, T. W. "Internal Revenue Service 990 Data: Fact or Fiction." *Nonprofit and Voluntary Sector Quarterly,* 1996, *25* (1), 40–52.

Grønbjerg, K. A. "Using NTEE To Classify Non-Profit Organizations: An Assessment of Human Service and Regional Applications." *Voluntas,* 1994, *5* (3), 301–328.

Kaple, D., Morris, L., Rivkin-Fish, Z., and DiMaggio, P. *Data on Arts Organizations: A Review and Needs Assessment, with Design Implications.* Princeton, N.J.: Center for Arts and Cultural Policy Studies, Princeton University, 1996.

Sales, G. *A Taxonomy of Human Services.* Los Angeles: Information and Referral Federation of Los Angeles County, 1991.

Smith, B. *The Use of Standard Industrial Classification (SIC) Codes to Classify Activities of Nonprofit, Tax-Exempt Organizations.* Working Paper No. 19. San Francisco: Institute for Nonprofit Organization Management, University of San Francisco, 1992.

Smith, D. H. "Four Sectors or Five? Retaining the Membership Sector." *Nonprofit and Voluntary Sector Quarterly,* 1991, *20* (2), 137–150.

Smith, D. H. "Public Benefit and Member Benefit Nonprofit, Voluntary Groups." *Nonprofit and Voluntary Sector Quarterly,* 1993, *22* (1), 53–68.