Establishing Data and Documentation Standards for Investigators who are Required to Archive Research Data

by Patrick T. Collins[1] , Project Director, National Data Archive on Child Abuse and Neglect

Introduction

This paper is about the approach that the National Data Archive on Child Abuse and Neglect has taken to improving the quality and consistency of our documentation. Of the many problems we have encounytered, including uncooperative investigators, dirty data files, and unusual file formats, poorly prepared or non-existent documentation has been the most difficult to handle. Since investigators were not required to archive their data with our Archive, we had to actively solicit contributors. Most researchers were unwilling to contribute their data and the ones who were willing had little or no resources to dedicate to the task. In short, we were in the position of having to accept whatever investigators were willing to provide. In many cases we received nothing more than an SPSS or raw data file and a copy of the instrument, leaving us with the daunting task of creating a user's guide from scratch. The task of preparing comprehensive documentation for these studies was so time consuming that were only able to process 2-3 datasets per year. While we appreciated the efforts of the investigators who chose to contribute data to the Archjive, it became clear that the only way the Archive could expand its holdings with any speed would be to improve the nature of the materials contributed by investigators. Since our Archive was funded by a federal agency with an active research program, we chose to work through that agency in order to establish data documentation standards for their research grantees. But before I describe this process in detail, let me tell you a little more about the Archive.

The National Data Archive on Child Abuse and Neglect

The National Data Archive on Child Abuse and Neglect has been in operation for approximately six years. During this time NDACAN has received all of its funding from the National center on child Abuse and Neglect (NCCAN) which is a division of the Administration for Children Youth and Families which in turn is a unit of the US Department of Health and Human Sercies. NCCAN is the federal agency witht the primary mission of responding to child abuse and neglect in the USA. One of NCCAN's many responsibilities is a field initiated research program which is funded at approximately $1.5 million per year. The Archive, which is funded through this program, works primarily with NCCAN's research grantees. We are in the final year of our second three-year award from NCCAN and will apply this spring for continued funding. Over its six years of operation, the Archive has been flat-funded at $150,000 per year, leaving us with approximately $100,000/year in direct funds. Most of these funds are used to support our 2.2 FTEs.

The Archive's primary mission is to acquire, process, preserve, and disseminate high quality datasets relevant to the study of child maltreatment. We have the secondary mission of networking and training child maltreatment workers. Toward this end, the Archive publishes a biannual newsletter, hosts a listserv with approxixmately 400 subscribers, and maintains a Gopher/FTP server. In many ways we have been more successful in achieving our secondary mission of networking and training researchers than our primary mission of acuirring datasets. creating networking and training opportunities is a fairly straightforward job and such services are eagerly consumed by researchers. Acquiring, processing, and disseminating high quality data is a far more complex task.

For these and other reasons, NDACAN began to advocate for mandatory data archiving for NCCAN research grantees. Simultaneously, we lobbied NCCAN to establish tecnical standards for their research grantees. Toward this end, Jane Powers and I co-authored, The Preparation of Data Sets for Analysis and Dissemination: Technical Guidelines for Machine-Readable Data, a manual which set forth standards for the preparation of research datasets and their associated documentation. We disseminated hundreds of copies of this manual to CNNAN's research grantees and to other child maltreatment researchers and we offered technical assistance to researchers willing to follow our guidelines. Unfortrunately very few researchers responded with interests. The situation changed however when, in their 1993 RFP, NCCAN announced that their research grantees would be expected to prepare datasets and documentation according to NCACAN's standrads. This generated quite a bit of interest and for the first time applicants began to contact us for technical assistance and copies of our manual.

The Archive's lobbying efforts came to fruition when, in their 1994 RFP, NCCAN set forth the requirement that applicants include in their proposal plans to prepare their data and documentation according to NDACAN's guidelines and to archive

their data with NDACAN upon the completion of their grant. As a result of this dramatic policy change, NDACAN will have the opportunity work with investigators from the beginning of their projects to ensure that data and documentation are prepared properly. all grantees will be provided with free technical assistance during the start-up phase of their studies and will receive a new publication entititled, Depositing Data with the National Data Archive on Child Abuse and Neglect: A Handbook for Investigators. The purpose of the handbook is to outline the investigators responsibilities and to provide a clear set of deliverables that must be submitted to NDACAN.

In some sense NCCAN's policy change took us by surprise. After years of lobbying, we were happily surprised to learn that NCCANdecided to require research grantees to archive their data. The way the requirement was implemented was that NCCAN reviews rated applicants' plans to prepare and archive data and documentation as one of the many criteria use to evaluate grant proposals. While it is not clear that this arrangement provides any method of enforcement, grantees are working under the assumption that they will be required to archive their data. While NCCAN set forth the requirement, the Archive is in the position of defining all of the nuts and bolts of the arrangement. Instead of reinvention the wheel, we have studied the data archiving programs of other federal agencies, such as the national Institute of Justice (NIJ) and the National Science Foundation (NSF). Our approach has been to build on the successes of these programs and make adjustments where necessary. Our goal is to build a program that meets the needs of NDACAN and NCCAN's research grantees.

In our experience of working with researchers, their greatest concern is having adequate time to publish their results of their study before the data are released to the public. In response, we have created a policy that will allow all investigators a two-year "grace period" after the termination of their grant which will allow them to publish the results of their study before the data are made available to the public. This is an area where our approach differs from that of NIJ. NIJ grantees must submit their data and documentation along with their final report at the termination of the award; grantees who fail to do so are not eligible for new NIJ funding. This policy has created a great deal of animosity among some NIJ grantees and there have been cases where a secondary data user published a study's findings before the principal investigator.

In other areas, we have closely followed NIJ's lead. For example, in determining the investigators' responsibilities and required deliverables we have essentially mimicked NIJ's requirements. Broadly defined, we see the investigators are responsible for, submitting data and sporting materials, responding to requests by NDACAN staff for additional or clarifying information, and reviewing and correcting draft materials prepared by the NDACAN staff.

NDACAN staff is responsible for preparing the data files in ready-to-use statistical file formats, preparing a user's guide that describes the project and data, reviewing the codebook for completeness and accuracy and augmenting the codebook as necessary, and making copies of the datasets archived available to the research community (for a small fee) and providing technical support to data users.

This new arrangement has the potential to solve the problem of inadequate and inconsistent documentation because we can specify exactly what materials the investigator must submit. Grantees will be provided with a clear set of deliverables as well as clear written guidelines for the preparation of those materials. Working with grantees early on in their projects in order to determine potential problem areas and needs for technical assistance will be integral to our approach.

While it will be several years before the first grantees are required to submit data under this arrangement we have established a tentative list of deliverables. These include:

(1) Data file(s).
(2) Description of data files
(3) Data collection instruments(s)
(4) References for data collection instruments
(5) Codebook or data dictionary
(6) Explanation of derived (computed) variables
(7) Final project report, project summary, or other description of the project
(8) Bibliography of publications pertaining to the data
(9) Printout of the first and last data records

The draft handbook that I have distributed contains some guidelines and specifications for the preparation of these materials. Our plan is to distill the most important guidelines in our technical standards manual and include them in the handbook. We want to keep the handbook as simple and free from jargon as possible. Once it is completed, we will stop distributing the technical standards manual because it is both out-dated and too technical for investigators. We are very interested in your

feedback and suggestions for the handbook so please read it if you have time. And let us know how we can improve it.

Once we receive these materials from the investigators we will create a comprehensive user's guide with the following format:

Project Overview
- Purpose of the Study
- Sampling/Selection Information
- Data Collection
- Instruments and Measures

Description of Machine-Readable Files
- List of Files
- Notes Regarding the Data Files

References
- References to publications from the dataset
- References to publications related to the dataset

Appendices
- Data Collection Instruments
- Codebook
- Sample Programs

So far reaction to the policy has been relatively positive. We presented our general plan to a large group of researchers at the annual NCCAN grantees meeting and most of the grantees were receptive. We are in the process of forming an advisory committee to handle cases where investigators have special needs relative to archiving (e.g., longitudinal studies). We are hopeful that NCCAN's data archiving policy will go a long way toward solving the problems I have described however, it will be several years before we know for sure. Clearly, the approach has limitations but we feel it is a step in the right direction.

1. Submitted for the IASSIST Conference held in Quebec, Canada. May 1995