

GUIDELINES FOR DATAFILE DOCUMENTATION

These guidelines are based on those used by the Inter-university Consortium for Political and Social Research. They should be used to document all datafiles deposited with the Institute of Governmental Affairs Library and Data Archive. Where possible, the documentation should be provided to Archive personnel in both printed and machine readable format (MS DOS Word or ASCII format), so that final codebooks with title pages, use restrictions, etc. can be easily generated.

The following items should be provided in the dataset documentation if appropriate (items marked with an * are required):

I. General Identification Elements*

A. Title of Survey, Study, Datafile, etc.*

The datafile should be given a descriptive title that reflects the objectives or content of the dataset and terminates with the year or years in which (or for which) the data were collected. If the study is a survey, the month(s) of data collection should be included. The title should be unique and assigned at the beginning of the data collection process. In the case of ongoing studies, once selected, this title should be adhered to thereafter in order to eliminate confusion.

B. Department/Agency*

The sponsoring agency or organization should be fully identified.

C. Principal Investigator(s)*

The individual(s) who are responsible for the study should be fully identified.

D. Institutions

Identify all institutions responsible for data collection, if different from the above. For example, as appropriate identify: institutions that drew up samples, institutions that administered questionnaires or interviews, institutions that made the data machine readable, the year that data were made machine readable, and the institution responsible for data analysis.

E. Contact Persons

Specify names and addresses of persons responsible for the survey (if other than the principal investigators). Individuals listed should be the appropriate persons to serve as resources regarding problems or questions raised by data users.

F. Edition Number and Date*

There may be multiple versions of a dataset as it is checked, revised, refined, expanded, etc. Datasets should be given edition numbers and dated to aid users and avoid confusion. The date given here should reflect when the version was created.

G. Archive and Statement of Availability*

This statement should indicate the terms of availability of the dataset to the public through the Institute of Governmental Affairs Library and Data Archive.

II. Disclaimer(s) and Other Notices to Users*

- A. It is important that original data collectors as well as data distributors not be held responsible for secondary use of the data and for interpretations and findings presented in print by secondary users. To that end, the following is a disclaimer that should be included in the documentation for a study:

The data (and tabulations) utilized in this (publication) were made available in part by the East Asian Business and Development Research Archive. The data for (title of study or dataset) were originally collected by (principal investigator(s) and/or agencies). Neither the original source or collectors of the data, nor the Research Archive bear any responsibility for the analyses or interpretations presented here.

- B. The following additional disclaimers or notifications recommended by University counsel should also be included:

This dataset is made freely available to the academic community for not-for-profit educational and research applications. Use of the data, either directly or indirectly, for profit is prohibited. In addition, the University of California makes no warranty respecting the accuracy of data furnished hereunder nor the results to be obtained from using such data for intended or other purposes.

- C. Users may be asked to submit copies of their publications, reports, working papers, etc. based on the use of the data to the Library and Data Archive. Users should be encouraged to inform the Data Archive of errors or discrepancies discovered while using the data. The following format may be adapted as needed for this purpose.

Individuals receiving and using these data are strongly urged to inform the Institute of Governmental Affairs Library and Data Archive of any errors or discrepancies that are discovered in the course of using these data. Users are particularly urged to contact the Archive about problems and difficulties which prevented effective and convenient utilization of the data. This information is necessary in order to improve the data and to facilitate more efficient and economical use of the data. Users are also asked to provide information as to significant subsets and special aggregations that are developed using these data. Finally, in order to provide agencies with essential information about the use of archival sources and to facilitate the exchange of information about research activities, each user is required to send two copies of each completed manuscript (or thesis abstract) to the Archive. This information should be addressed to Shelagh M. Mackay, Coordinator, Library and Data Archive, Institute of Governmental Affairs, University of California, One Shields Avenue, Davis, CA 95616-8617.

III. Bibliographic Citation*

Users of data are obligated to cite the data upon which their publications or reports are based. The recommended citation format should be provided in the documentation and should include the following elements:

- A. Principal investigator and/or corporate authors
- B. Title of the study followed by a bracketed reference [machine readable datafile]
- C. Organization which produced the data followed by a bracketed reference "[producer]" and the year of production (if known).*

- D. Distributor followed by a bracketed reference "[distributor]" and the date the data became publicly available (if known).*

IV. Methodology*

This section provides the substantive information required to evaluate and use the study data. It serves as a valuable written history of the study or dataset design, conceptualization, and implementation. As available and appropriate, it may also provide an appraisal of the data which discusses the measurement and validity of the data, and a list of publications and reports based upon the data. Specific elements may include the following.

- A. Study Objectives

A precise and detailed statement of the problem to be examined and goals to be met by data collection should be included here.

- B. Abstract of Dataset*

A brief description or abstract of the dataset content should be provided, including the dates of data collection or coverage.

- C. Relation to Other Surveys and Programs.

Information should be provided on other related studies or programs, as appropriate and available.

- D. Study Design*

This section should discuss the design employed in the dataset and the procedures used. This may include such information as operational procedures and sample design. Categories to be covered might include 1) universe or sample population, 2) sample design, 3) types of instruments used, how they were administered and to whom they were administered, 4) description of pilot study or pretest, if used, 5) appraisal of the data (such as sampling error, response variance, nonresponse rate and testing for bias, interviewer and response bias, confidence levels, question bias).

- E. Publications List*

This section can include a bibliography of publications and reports based on the data.

V. Technical Information*

This section documents the extent to which data were cleaned and checked. It also provides information about the structure of the file(s). Categories to be covered (as applicable) in this section are outlined below.

- A. Data Checks

This section should outline data checking procedures such as whether or not data sorted and merge-checked, checked for "wild" codes or data entry errors, or checked for inconsistent responses.

- B. Missing Data*

This section should provide information on how missing data or nonresponses were handled. For example, is there a missing data code, is the variable left blank, etc.

- C. Structure of Data*

This section should state basic information about the data structure. For example, is the dataset in card-image or logical record format on computer tape; is it a microcomputer file; if so, what is the

data format? This section should also state how many respondents or cases are in the file; the usage of unique identification numbers and how they are sorted.

D. Number of Files*

In this section, if more than one separate file exists for the study, provide the basic information noted above for each file.

E. Weights

If the data are weighted, describe the weights, the location of the weight variable, and when the data should be weighted.

VI. List of Variables

This section should provide a list of variables provided in the dataset in summary form. For a survey, the questionnaire might be provided here.

VII. Technical Documentation*

This section should provide the complete technical and substantive description of each question or variable as well as the location of the variable in the record (on disk, tape, etc.). It will serve as the codebook to the dataset. For a survey, this codebook can be a retyped or machine-readable version of the questionnaire containing all the questions in the order they were administered, along with their associated codes and technical information. The following elements should be included in this section as appropriate.

A. Variable Names and Numbers*

Number the variables in the order in which they appear in the file. For each variable, give the variable (or field) name. When data are prepared for use with certain types of software such as SPSS or SAS other optional items may be provided.

B. Variable Location

For a dataset stored on tape, tape location information must also be provided. For each variable or question, give information such as deck or column number (for card image data), or precise tape location (for a dataset in logical record format).

C. Question Text or Variable Description*

For a survey, the complete question text should be given. For other types of data, provide a brief description of the variable.

D. Explanatory Text

Provide any background information on the variable in this section. For a survey, this might include interviewer instructions. For other types of data, appropriate elements will include data source or derivation.

E. Code Categories*

Where data are coded, provide a complete explanation of the coding scheme and any abbreviations employed.

F. Missing Data*

Any missing data for each variable should be carefully explained. For a survey, if questions were administered to only a subset of the population that subset needs to be described. For other types of data, document how fields were coded when data were unavailable.

G. Frequencies, Percentages, Summary Statistics

Frequencies, percentages or summary statistics may be incorporated here as appropriate.

H. Newly Created Variables*

During the data analysis process, new variables are frequently created by combining responses across several variables. If such variables were created, document the procedures used to create them.

VIII. File Specifications and Data Format*

This section should provide the following information on file specifications and data format. For a study on tape, this will include such information as whether the file is in 80-character card-images or logical records of character strings. For logical records, is the data in BCD, EBCDIC, or ASCII format, are the data rectangular or flat file, etc. For studies on floppy diskette, what is the file format (dBASE, Lotus worksheet, SPSS), file size (in bytes and the number of records), etc.

STUDY IDENTIFICATION: S-CAPS-8905

TITLE: The California Poll, December 1989

PRODUCER #: 8905

PRINCIPAL INVESTIGATOR: The Field Institute

SERIES: California Polls

DISTRIBUTOR: The Field Institute

DATE OF DATA COLLECTION: December 4-7 and December 8-13, 1989

UNIVERSE: Representative cross section of adults

GEOGRAPHY: California

CONTENTS OF FILE: Issues Relating to Senator Cranston: Heard of Charles Keating; think Cranston did anything wrong in his dealing with Keating; why was Cranston wrong (list); satisfied with Cranston public response to Keating questions; Opinion of Cranston investigation; vote for Cranston if ran again. Registration and Party Affiliation: Registered to vote in own precinct, elsewhere or not at all; party affiliation. Californians: Born in CA; if no, where born; migration history; years residency in CA; rate CA as place to live; rate CA with others; quality of CA's life; description of Californians: tolerant, traditional, culturally diverse, fun loving, trendy, conservative, self centered, fortunate, ethnically diverse, friendly, "laid back", God fearing, sophisticated, liberal, innovative, old fashioned, enterprising, generous, resourceful, polite, self indulgent, aggressive, money oriented, family oriented, status conscious, ethnical. Capital Punishment and the Court System: favor death sentence; opinion: better to let some guilty free than convict an innocent person, criminals should be punished harshly, a person brought to trial is probably guilty, criminals should be considered for mercy, illegal evidence should not be allowed, prisons should punish rather than rehabilitate criminals, insanity plea is a loophole, harsher treatment of criminals is not the solution, favor death penalty. Hypothetical 'ur decisions: opinion on death penalty interferes with ability to act fairly as juror; should death penalty be imposed on juveniles; should death penalty be imposed on the mentally retarded; more or less likely to vote for the death penalty when: the murder was not premeditated; the murder was extremely brutal; the murder was committed under the influence of drugs/alcohol; there was more than one murder involved; the murder was the only crime ever committed, the murder included sexual assault, the murder was committed under extreme mental or emotional disturbance; until the time of the crime, the person was good; convicted person was over the age of 30; convicted person had committed prior felony; convicted person came from a background of extreme poverty; convicted person had been seriously abused as a child; convicted person would be well-behaved inmate in prison; convicted person had previously been in prison but had received no help; convicted person did not express any remorse or regret; convicted person has a loving family; Opinion: death penalty prevents murders, Bible supports the death penalty, death penalty is cheaper than life in prison, should the decision to impose the death penalty depend on the background and characteristics of the convicted person, retribution is a sufficient argument for the death penalty, innocent people are too often executed, minorities are more likely to receive the death penalty, life in prison without possibility of parole does not guarantee that a prisoner will not be released from prison, death preferable to life in prison. Crime: danger from crime in city, neighborhood has increased in the last year; house broken into in the past 12 months, money or property stolen, car stolen, home or car vandalized, respondent assaulted: did not go because unsafe part of town. Gang Activity: Effect of street gang activity; seriousness of gang activity: Proposals: put money into youth job programs; put money into counseling and social services; put money into police force; pass tough laws against gang activity; favor metal detectors in schools. Domestic Violence: Punishment: Husband-Wife argument, physical discipline, verbal threat by parent; seriousness of verbal/physical damage; was respondent victim of parental violence; frequency of parental violence: victim of violence by spouse: frequency of spouse violence: frequency of arguments between spouse. Political Opinions: Opinion: Senator Cranston; rate Cranston: performance, honesty, investigation results, reelection, resignation, effectiveness: rate California U.S. Senate: Pete Wilson better as Senator or Governor. Political Recognition: Name Senators: Cranston, de Concini, Glenn, McCain, Reigle, others. Background of Respondent: Marital status, age, education, political ideology, party identification/strength, religion, income, Hispanic descent, race, own or rent, gender, county.

FILE DESCRIPTION: 1 data file (3,027 records) + 1 codebook file (5,016 records) + 1 SPSS control cards file (378 records) + associated documentation
RECORDS PER CASE: 3 records, 80 characters in length
OF VARIABLES: 151 STATUS: UC use only

As outlined in Carolyn L. Geda and John D. Peine. *Data Preparation Manual*. Ann Arbor, Mich.: Inter-university Consortium for Political and Social Research, January. 1980.