Technical Standards for Machine-Readable Data

Supplied to the Law Enforcement Assistance Administration

U.S. Department of Justice

Law Enforcement Assistance Administration

National Criminal Justice Information and Statistics Service

National Criminal Justice Information and Statistics Service Reports

Single copies are available at no charge from the National Criminal Justice Reference Service, Box 6000, Rockville, Md. 20850. Multiple copies are for sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402

Applications of the National Crime Survey Victimization and Attitude Data:

Public Opinion About Crime: The Attitudes of Victims and Nonvictims in Selected Cities, NCJ-41336

Local Victim Surveys: A review of the Issues, NCJ-39973

The Police and Public Opinion: An Analysis of Victimization and Attitude Data from 13 American Cities, NCJ-42018

An Introduction to the National Crime Survey, NCJ-43732

Compensating Victims of Violent Crime: Potential Costs and Coverage of a National Program, NCJ-43387

Crime Against Persons in Urban, Suburban, and Rural Areas: A Comparative Analysis of Victimization Rates, NCJ-53551

Victimization Surveys:

Criminal Victimization in the United States (annual)
- A Comparison of 1976 and 1977 Findings, Advance Report, NCJ-52993
- A Comparison of 1975 and 1976 Findings, NCJ-44132
- A Comparison of 1974 and 1975 Findings, NCJ-39548
- A Comparison of 1973 and 1974 Findings, NCJ-34391
- 1976 (final report), NCJ-49543
- 1975, NCJ-44593
- 1974, NCJ-39467
- 1973, NCJ-34732

The Cost of Negligence: Losses from Preventable Burglaries, NCJ-53527

Criminal Victimization Surveys in

Boston, NCJ-34818	**New Orleans,** NCJ-34825
Buffalo, NCJ-34820	**Oakland,** NCJ-34826
Cincinnati, NCJ-34819	**Pittsburgh,** NCJ-34827
Houston, NCJ-34821	**San Diego,** NCJ-34828
Miami, NCJ-34822	**San Francisco,** NCJ-34829
Milwaukee, NCJ-34823	**Washington, D.C.** NCJ-34830
Minneapolis, NCJ-34824	(final report, 13 vols.)

Criminal Victimization Surveys in 13 American Cities (summary report, 1 vol.), NCJ-18471

Public Attitudes About Crime:

Boston, NCJ-46235	**New Orleans,** NCJ-46242
Buffalo, NCJ-46236	**Oakland,** NCJ-46243
Cincinnati, NCJ-46237	**Pittsburgh,** NCJ-46244
Houston, NCJ-46238	**San Diego,** NCJ-46245
Miami, NCJ-46239	**San Francisco,** NCJ-46246
Milwaukee, NCJ-46240	**Washington, D.C.** NCJ-46247
Minneapolis, NCJ-46241	(final report, 13 vols)

Criminal Victimization Surveys in Chicago, Detroit, Los Angeles, New York, and Philadelphia: A Comparison of 1972 and 1974 Findings, NCJ-36360

Criminal Victimization Surveys in the Nation's Five Largest Cities: National Crime Panel Surveys in Chicago, Detroit, Los Angeles, New York, and Philadelphia, 1972, NCJ-16909

Criminal Victimization Surveys in Eight American Cities: A Comparison of 1971/72 and 1974/75 Findings—National Crime Surveys in Atlanta, Baltimore, Cleveland, Dallas, Denver, Newark, Portland, and St. Louis, NCJ-36361

Crimes and Victims: A Report on the Dayton-San Jose Pilot Survey of Victimization, NCJ 013314

National Prisoner Statistics:

Capital Punishment (annual)
- 1977 (final report), NCJ-49657

Prisoners in State and Federal Institutions (annual)
- December 31, 1977 (final report), NCJ-52701

Census of State Correctional Facilities, 1974: Advance Report, NCJ-25642

Survey of Inmates of State Correctional Facilities, 1974: Advance Report, NCJ-34267

Census of Prisoners in State Correctional Facilities, 1973, NCJ-34729

Uniform Parole Reports:

Parole in the United States: 1976 and 1977, NCJ-49702

Census of Jails and Survey of Jail Inmates, 1978: Preliminary Report, NCJ-55172

The Nation's Jails: A report on the census of jails from the 1972 Survey of Inmates of Local Jails, NCJ-19067

Survey of Inmates of Local Jails 1972: Advance Report, NCJ-13313

Children in Custody: Juvenile Detention and Correctional Facility Census
- Advance Report, 1975 census, NCJ-43528
- Advance Report, 1974 census, NCJ-38820
- Final Report, 1973 census, NCJ-44777
- Final Report, 1971 census, NCJ-13403

Myths and Realities About Crime: A Nontechnical Presentation of Selected Information from the National Prisoner Statistics Program and the National Crime Survey, NCJ-46249

State Court Caseload Statistics:
- The State of the Art, NCJ-46934
- Advance Annual Report, 1975, NCJ-51884
- Annual Report, 1975, NCJ-51885

National Survey of Court Organization:
- 1977 Supplement to State Judicial Systems, NCJ-40022
- 1975 Supplement to State Judicial Systems, NCJ 29433
- 1971 (full report), NCJ-11427

State and Local Probation and Parole Systems, NCJ-41335

State and Local Prosecution and Civil Attorney Systems, NCJ-41334

Trends in Expenditure and Employment Data for the Criminal Justice System, 1971-76 (annual), NCJ-45685

Expenditure and Employment Data for the Criminal Justice System (annual)
- 1977 final report, NCJ-53206

Criminal Justice Agencies in Region
- **1:** Conn., Maine, Mass., N.H., R.I., Vt., NCJ-17930
- **2:** N.J., N.Y., NCJ-17931
- **3:** Del., D.C., Md., Pa., Va., W. Va., NCJ-17932
- **4:** Ala., Ga., Fla., Ky., Miss., N.C., S.C., Tenn., NCJ-17933
- **5:** Ill., Ind., Mich., Minn., Ohio, Wis., NCJ-17934
- **6:** Ark., La., N. Mex., Okla., Tex., NCJ-17935
- **7:** Iowa, Kans., Mo., Nebr., NCJ-17936
- **8:** Colo., Mont., N. Dak., S. Dak., Utah, Wyo., NCJ-17937
- **9:** Ariz., Calif., Hawaii, Nev., NCJ-15151
- **10:** Alaska, Idaho, Oreg., Wash., NCJ-17938

Dictionary of Criminal Justice Data Terminology: Terms and Definitions Proposed for Interstate and National Data Collection and Exchange, NCJ-36747

Program Plan for Statistics, 1977-81, NCJ-37811

Utilization of Criminal Justice Statistics Project:

Sourcebook of Criminal Justice Statistics 1977 (annual), NCJ-38821

Public Opinion Regarding Crime, Criminal Justice, and Related Topics, NCJ-17419

New Directions in Processing of Juvenile Offenders: The Denver Model, NCJ-17420

Who Gets Detained? An Empirical Analysis of the Pre-Adjudicatory Detention of Juveniles in Denver, NCJ-17417

Juvenile Dispositions: Social and Legal Factors Related to the Processing of Denver Delinquency Cases, NCJ-17418

Offender-Based Transaction Statistics: New Directions in Data Collection and Reporting, NCJ-29645

Sentencing of California Felony Offenders, NCJ-29646

The Judicial Processing of Assault and Burglary Offenders in Selected California Counties, NCJ-29644

Pre-Adjudicatory Detention in Three Juvenile Courts, NCJ-34730

Delinquency Dispositions: An Empirical Analysis of Processing Decisions in Three Juvenile Courts, NCJ-34734

The Patterns and Distribution of Assault Incident Characteristics Among Social Areas, NCJ-40025

Patterns of Robbery Characteristics and Their Occurrence Among Social Areas, NCJ-40026

Crime-Specific Analysis:
- The Characteristics of Burglary Incidents, NCJ-42093
- An Empirical Examination of Burglary Offender Characteristics, NCJ-43131
- An Empirical Examination of Burglary Offenders and Offense Characteristics, NCJ-42476

Sources of National Criminal Justice Statistics: An Annotated Bibliography, NCJ-45006

Federal Criminal Sentencing: Perspectives of Analysis and a Design for Research, NCJ-33683

Variations in Federal Criminal Sentences: A Statistical Assessment at the National Level, NCJ-33684

Federal Sentencing Patterns: A Study of Geographical Variations, NCJ-33685

Predicting Sentences in Federal Courts: The Feasibility of a National Sentencing Policy, NCJ-33686

Technical Standards for Machine-Readable Data Supplied To the Law Enforcement Assistance Administration

Report Number (SD-T-2)

This report was supported by Grant No. 78-SS-AX-0028

Awarded to The Bureau of Social Science Research, Washington, D.C., by the Statistics Division, National Criminal Justice Information and Statistics Service, Law Enforcement Assistance Administration, U.S. Department of Justice, under the Omnibus Crime Control and Safe Streets Act of 1968, as amended. The project which produced this report was directed for the Bureau of Social Science Research by Richard C. Roistacher and monitored for LEAA by Marianne W. Zawitz.

U. S. Department of Justice
Law Enforcement Assistance Administration

Henry S. Dogin, Administrator

Homer F. Broome, Jr., Deputy Administrator
for Administration

Benjamin H. Renshaw, Acting Assistant Administrator
National Criminal Justice Information
and Statistics Service

Charles R. Kindermann, Acting Director
Statistics Division

TABLE OF CONTENTS

Technical Standards for Machine-Readable Data Supplied To the Law Enforcement Assistance Administration

This standard sets forth technical requirements for data supplied to the Law Enforcement Assistance Administration under grants, contracts, and interagency agreements. All machine-readable data produced for LEAA which relate to issues of national importance or interest, or which are national in scope of coverage must meet these technical standards and must be submitted to the National Criminal Justice Data Archive.

LEAA has established the National Criminal Justice Data Archive at the University of Michigan. The Archive is designed to promote the analysis of criminal justice data by making inexpensive data tapes available in an easy-to-use form. LEAA's intent when establishing the Archive was not only to make available data from its own statistical series, but also to provide all important machine-readable data relevant to national criminal justice issues. Many such data files are developed under LEAA grants and contracts.

LEAA has established these technical standards to ensure the quality and utility of data submitted to the Archive. Although most of these standards are simply good data processing practice, some of the requirements are specific to the needs and facilities of LEAA, and are noted as such. Additional information about these standards and the Archive may be obtained from the Statistics Division of the National Criminal Justice Information and Statistics Service (NCJISS).

NCJISS, Statistics Division, will provide assistance in determining the degree of compliance required and in monitoring compliance. Assistance in meeting the technical standards is available to grantees from the University of Michigan. Requests for either type of assistance should be referred to NCJISS, Statistics Division.

Tape Recording Standards

Data supplied to LEAA should be on 9-track magnetic tape written at a density of 800, 1600, or 6250 bytes (characters) per inch. Since more reliable encoding

methods are used at the higher densities, 6250 BPI tapes are preferred to 1600 BPI tapes, which are preferred to 800 BPI tapes. The tape should be labelled with IBM or ANSI volume and file labels. If 9-track equipment is not available, then files may be submitted on 7-track magnetic tape. Either fixed or variable length records are acceptable, and files should be blocked to a length which will provide for efficient handling of the data. Physical record lengths of from 80 to 32,767 characters are acceptable.[1,2] Volume and file-naming conventions should be agreed on by LEAA and the data supplier before tapes are written. While it is impractical to devise a set of file-naming conventions which are universally applicable, it will often be in the interest of LEAA and its data suppliers to establish naming conventions for a particular set of data files.

Data Types

All machine-readable data records supplied to LEAA will consist entirely of alphanumeric, EBCDIC, or ASCII characters. (Length fields in binary format which are part of the IBM variable-length record are not included in this restriction.)

For the purpose of this standard, data items will be considered either computational or noncomputational. Noncomputational items contain information such as names

---

[1] Data on magnetic tape are divided into physical and logical records. A logical record cont[illegible]s data from a single "unit" (e.g., defendant, gas bi[illegible], library catalog card, etc.). Data processing is often more efficient if several logical records are written end-to-end onto the tape to form a single "physical" record. This process is called "blocking", and a physical record is sometimes called a "block".

[2] This requirement is specific to the National Criminal Justice Data Archive, whose computer can handle extremely long physical records. In general, data producers and users should consult with the technical staff of their computing facility to determine the facility's preferred size for physical records. The American National Standards Institute recommends a maximum physical record size of 2400 characters, which is considerably less efficient than the maximum permitted by this standard.

and labels, that are never used in any arithmetic or numerical operation. Noncomputational data items may contain any EBCDIC or ASCII printing characters. Noncomputational items will be left justified and padded to the right with alphabetic blanks.

Computational items contain numeric information that is designed to be used in computations. Computational data items may contain only the characters Ø-9, ".", "+", and "-". The only exception to this is where D- or E-format floating-point data are represented.[3] Wherever possible, data should be represented in integer format with implicit decimal places noted in the documentation. Computational data items should be right justified and padded to the left with zeros or blanks. If a computational field is signed, the sign must immediately precede the left-most numeric character. Computational fields must contain at least one numeric character. In particular, computational fields consisting only of a sign or of blanks are not permissible. Fields containing only a signed zero are not acceptable, since some computers cannot represent a signed zero.

Missing data. All data items for which there may be missing data must have an explicit missing data value or values. In particular, it will never be assumed that a value of zero or a field consisting entirely of blanks indicates missing data. Missing data values may be indicated in the documentation as a list or a range of values, and may include zero, but may not include values of -Ø, a blank field, or a field consisting only of a sign.

Missing data values must occur in the same field as the variable to which they refer. If an alternate value is to be used in place of a missing data value, the base

---

[3] Floating-point data are used to represent very large or very small numbers. The number, 2,625,ØØØ,ØØØ is written in floating-point notation as "2.625 E+Ø9". The "E+Ø9" is called the exponent, and indicates that the decimal point is to be shifted nine places to the right. D-format floating-point notation is similar, but uses a "D" rather than an "E" to mark the exponent. The number "Ø.ØØØØØØØØØØØ465" is written in floating-point notation as "4.65 E-12". The "D" indicates that the number was computed with double precision arithmetic, which allows more significant digits to be represented.

item must carry an appropriate missing data code, while the alternate value will be shown in a separate data item which has been declared for that purpose. In no case will the alternate value be carried in the base variable, with an explanation code in another variable. The rationale for this standard is that the meaning of a variable should be determinable without reference to a second variable. If the true value of a variable has been suppressed or modified, then the value of the data item should indicate such suppression or modification. If an alternate value is to be offered in such cases, the appropriate data item in the record can then be read if the analyst so wishes.

File Organization

Data items. Records transmitted to LEAA should contain no undocumented or irrelevant fields. The width of a data item should be sufficient to accommodate the entire range of variation which may be expected of the item, but should not be excessive. Conversely, fields need be no wider than is required to accommodate the maximum expected value of a variable. Thus, the data item "number of offenders" for a multiple victimization incident probably need not be larger than two characters, and certainly no larger than three.

In any case, the maximum allowable field width for an integer variable on tapes submitted to LEAA is nine characters. Larger numbers should be represented by supplying a scaling factor[4] in the documentation, or by the use of floating-point format. A noncomputational data item may be up to 32,767 characters long.

Record type identification. Each record in a file containing more than one record type, e.g., a hierarchical file, will carry a data item which identifies the type of record. The record identifier will be the first data item in each record. Even where each record in a file has a different length, a data item identifying the type of record will be included. A new value of this identifier will be used whenever there is a significant change in any

---

[4] A scaling factor allows fewer characters to be used in representing a data item. For instance, $50,000,000 can be represented as "50" if the data item is defined as "Dollars in millions."

of the procedures used to generate that type of record. Such changes include changes not only in record layout but also in instrument design, data collection, and coding. The need for a new record type is obvious when coding instructions or code values are changed. However, even when such changes consist only of the addition of new coding categories to existing data items, a new record type should be produced. Otherwise, analysts may not properly interpret the absence of a particular response.

For example, several minor changes were made in the coding categories and format of the incident record of the National Crime Survey Victimization file. In order to interpret a "type I" record properly, it is necessary to know what year the data in the record represent. The record type and collection period should be used to generate a set of identifiers for incident records, identifying the particular format used.

Record identification items. Each record in a file submitted to LEAA must carry an identification number which is unique to the data file. If no existing data item will suffice as a unique identifier, then a sequence number should be assigned by the data supplier. Where a file contains more than one type of record, each record shall carry a data item identifying the type of record, as well as a unique sequence number. Where the records in a file represent a hierarchy or tree, a record will carry identification sufficient uniquely to identify it and its position in the hierarchy. In particular, it should not be necessary to infer the location of a record in a hierarchy solely from its position in the file.

For example, consider a file consisting of household, person, and incident records. Each record in the file must begin with the same four data items: A record type indicator, a household identifier, a person identifier, and an incident identifier. An incident record will carry a type and an incident identifier, as well as the identifiers of the person and household to which it belongs. A person record will carry a type identifier, the identifier of the household to which it belongs, its own identifier, and a dummy incident identifier. Record identifiers should be positive integers. Dummy identifiers should be fields of zeros. Blank record identification data items are not permitted.

In general, identifiers of lower level records need be unique only within level, since concatenating identification variables generates a unique identifier.[5] In some files, there will be more than one type of record at the same level of the hierarchy. Questions as to whether sequence numbers should be unique within record type or within levels should be resolved by agreement between the data supplier and LEAA before the file is generated.

The rationale underlying the assignment of a unique identifier to each record is that users of a data file should be able to perform arbitrary sorts and to subset the file without requiring the use of any facilities other than a sort program and a file-copying utility.

Standardization Of Data Codes

Where possible, standard data codes should be used. While there are no universal standards, it is often possible to find a widely used set of codes which is appropriate. Geographic coding can be done either with FIPS (Federal Information Processing Standards) codes or with Census Bureau codes. The choice of which "standard" coding scheme is used is not so important as that ad hoc coding schemes should not be used when equally good coding schemes are already in use. Data producers who have created what they believe to be generally useful coding schemes should indicate this fact to LEAA. LEAA will attempt to disseminate such information.

Documentation

Each tape submitted to LEAA must be accompanied by documentation giving the physical characteristics of the volume and files, as well as the logical composition of each type of record. Wherever possible, the documentation should be in machine-readable form and supplied both in

---

[5]To concatenate identifiers is to string them into a single identifier. Consider a file of households with people in them. Suppose that each household has a three-digit identifier which is unique within the file, and each person has a two-digit identifier which is unique within his or her household. Then person number 4 in household number 207 is uniquely identified as person number 20704.

hard copy and as a file on the tape. Machine-readable documentation is preferred because it prevents the separation of documentation from data and because the physical quality of the documentation will not be degraded by repeated copying.

Machine-readable documentation should be in printer image. If a document processing program is used to format the source text, it is requested that the source text of the documentation be included as well. (This standard makes no specification of any particular document processing program, nor does it require that such a program be used at all.)

Minimal documentation. The minimal documentation of a data file consists of a tape volume table of contents, a character and octal or hexadecimal dump of a sample of records of each file, and a minimal codebook. The name, title, and affiliation of the principal investigator responsible for collection of the data submitted, the source of funding (including the grant number, if any) must be included in the documentation. Where appropriate, data files should be accompanied by copies of the original collection instruments, including survey questionnaires and interview schedules. Copies of editing and coding instructions used in creating the data file should also be included.

Tape table of contents. The tape volume table of contents listing should include all information from the volume label and from the file labels. Information should be in an easy-to-read form, rather than a dump of the text of the labels. The following figure is an example of a tape volume table of contents listing.

If possible, the tape table of contents should be produced by a program which also verifies the readability of the tape. If no tape listing program is available, then the tape table of contents should be produced manually, using information from the job which produced the tape or a dump of the tape. (Since tape listing programs are becoming increasingly available, it is suggested that data centers without such programs attempt to acquire them.)

Minimal codebook. The codebook included with the file must contain at least the following information for each data item:

1. A reference number.

2. An unambiguous name for the item.

3. A textual description of the item, or the text of the question, if from a questionnaire.

4. The starting location, width, location of implicit decimal point, or scale factor.

5. Missing data codes and their meanings.

6. The mode in which the data item is represented, i. e., numeric character, alphanumeric string, floating-point binary, etc.

The codebook must also contain a list of the valid values for categorical items, and valid ranges for continuous items. Missing data codes must be documented in the same fashion as other values, and not left implicit.

Frequency tables. A frequency distribution for each categorical data item must accompany each file submitted to LEAA. The mean, standard deviation, range, and number of cases of continuous data items should also accompany the file. Values which fall outside of those defined in the codebook should be annotated if they cannot be corrected.

A tape volume table of contents

TAPE NAME = *RSW009* 24 MAY 1977 12:20:06

IBM-LABELED 6250-BPI 9TP VOLUME=RSW009 OWNER=CAC,U-ILL RACK#=C4524
LP=ON BLK=ON RING=OUT DTCHK=ON RETRY=10

FILE #	DATA SET NAME	BLOCK COUNT	RECORD COUNT	TAPELTH (FEET)	RECORD FORMAT	BLOCK LTH AV.	BLOCK LTH MAX.	CREATED DD MMM YY	EXPIRES DD MMM YY	USER I.D.	BATCH RECEIPT#
1	NCS73.NAT.CQ1	1233	125858	283.07	VB(15250,305)	15141	15250	23 MAR'77		SGDA	
2	NCS73.NAT.CQ2	1234	126442	283.17	VB(15250,305)	15135	15250	23 MAR'77		SGDA	
3	NCS73.NAT.CQ3	1187	123279	272.60	VB(15250,305)	15146	15250	23 MAR'77		SGDA	691127
4	NCS73.NAT.CQ4	1212	125436	278.21	VB(15250,305)	15139	15250	23 MAR'77		SGDA	691127
5	NCS74.NAT.CQ1	1086	112135	249.39	VB(15250,305)	15139	15250	23 MAR'77		SGDA	691127
6	NCS74.NAT.CQ2	1083	112086	248.84	VB(15250,305)	15148	15250	23 MAR'77		SGDA	691127
7	NCS74.NAT.CQ3	1072	111090	246.18	VB(15250,305)	15138	15250	23 MAR'77		SGDA	691127
8	NCS74.NAT.CQ4	1100	113455	252.62	VB(15250,305)	15141	15250	23 MAR'77		SGDA	691127

TOTAL TAPE LENGTH = 2114.07 FEET

<*><*><*> END OF TAPE <*><*><*>

*U.S. GOVERNMENT PRINTING OFFICE: 1979 O—281-387/19

END