Topically-focused Data Archives:

A New Paradigm for the Codification of Social Science Research

by Josefina J. Card
President, Sociometrics Corporation
3191 Cowper Street
Palo Alto, California 94306
Tel. (415) 321-7846

The "information explosion" has become a distinguishing feature of modern science. Both the published scientific literature and its supporting data files continue to grow at unprecedented rates. More than ever, it has become important that efficient ways be found to store available information on a given topic and then retrieve relevant portions of that information as they are required. In the 1970's enormous progress was made in the development of procedures to store and retrieve bibliographic information. The DIALOG, ERIC, and MEDLARS databases are but a small sample of the growing number of computerized bibliographic databases available to social scientists. Less significant progress has been made in the development of analogous procedures to store, catalog, and retrieve elements common to the numeric (or raw data) information underlying the published studies. Enormous productivity and cost savings could result from such development. For little additional cost relative to the data collection costs already incurred, a substantively-focused data archive with indexing capabilities could: accelerate the growth and dissemination of scientific knowledge about a topic of contemporary interest; encourage corroboration and replication of newly reported findings; provide policymakers and practitioners with a larger scientific base on which to build their work; and stimulate investigations by new investigators without access to the substantial funds required for new data collection. This paper introduces the Data Archive on Adolescent Pregnancy and Pregnancy Prevention (DAAPPP), to illustrate features of an emerging information resource: the special-purpose social science data archive.

The accumulation of knowledge about human reproduction, coupled with the development of relatively safe, effective, and inexpensive contraceptive methods, has made it possible for human beings to seize control of their biological destinies, and to plan the size and spacing of their families. Differences continue to persist,

however, in the degree to which various groups of people have been able to avoid unplanned and unwanted pregnancies. Rates of unplanned and unwanted pregnancies are higher in the developing than in the developed world. In a given country, young unmarrieds and the socially and economically disadvantaged have generally been found to be more vulnerable. The rate of out-of-wedlock pregnancy and childbearing among U.S. teenagers is among the highest in the world. DAAPPP was established by the U.S. Office of Population Affairs of the Office of the Assistant Secretary for Health to encourage the conduct and dissemination of research on these important social issues.

DAAPPP identifies, selects, acquires, and archives the most valuable databases dealing with U.S. adolescent fertility and U.S. family planning. Database *identification* refers to the systematic identification of all machine-readable data sets capable of addressing these topics. Database *selection* refers to the selection, from the identified universe, of the most outstanding data sets to include in the collection. Technical quality, substantive scope, and policy relevance are considered simultaneously by a National Advisory Panel of scientists in making selection decisions. Database *acquisition* refers to obtaining selected data sets from their holders. The raw data, the documentation, and completed reports and publications are all acquired. Database *archiving* refers to the processing and documentation of acquired data sets by archive staff, so that standardized, easy-to-use products are produced and disseminated. DAAPPP then makes its data and documentation publicly available, for the cost of reproduction. The following products are now publicly available for each of the 45 data sets currently in DAAPPP (see Table 1):

- A computer tape for use with mainframe computers with two machine-readable files:
  - the raw data;
  - SPSS-X program statements to convert the raw data to an SPSS-X system file (SPSS is an acronym for the widely-used Statistical Package for the Social Sciences).
- Floppy diskette(s) for use with microcomputers (in either 360-kilobyte or 1.2-megabyte format), with two machine-readable files:
  - the raw data;
  - SPSS/PC program statements to convert the raw data to an SPSS/PC microcomputer system file.
- A printed and bound user's guide, with five standard sections:
  - an overview of the original purpose for which the data were collected, and a description of the file's processing history;
  - a description of the machine-readable files available for the data set;
  - a categorization of the variables included in the data set by their topic and type, followed by a listing of all variables, sorted by topic and type;
  - a report on the completeness and quality of the data;
  - a bibliography of representative publications based on the data set.
- The codebook and instrument from the original investigation, where available.

Users of statistical package programs know that part of the routine procedure in the development of system files for analysis is the assignment of names and labels to variables in the file. We use the output of this routine procedure—lists of variable names and values—in an innovative way to give the archive indexing capabilities.

Each variable in DAAPPP is given an eight-character name for use with SPSS-X or SPSS/PC, to standardize variable names across all files, and to provide the user with quick reference to certain useful information about the

variable. Characters 1–2 encode the variable's TOPIC, the main subject matter of the variable. Character 3 encodes the variable's TYPE, further classifying the subject matter into one of the many variable types commonly used by social scientists. Characters 4–5 are a reference to the DATA SET ID, indicating the original source of the data. Characters 6–8 contain the VARIABLE SEQUENCE NUMBER, indicating the sequential position of the variable within the source data set. Table 2 contains the list of TOPICS, Table 3 the list of TYPES. Definitions of each topic and type have been developed that provide an inter–rater categorization reliability of over 90%. The list of DATA SET IDs is shown in Table 1. Each list can be altered easily to suit archives focused on other substantive topics. The variable naming scheme encodes information both on what each variable has in common with other variables in the archive (its TOPIC and TYPE), and what is unique to the variable (its SEQUENCE NUMBER with a given DATA SET ID).

DAAPPP staff members have written a simple computer program that uses the TOPIC and TYPE characters of the variable names as input, to produce a matrix that depicts, at a glance, the topical emphasis of each data set. For example, DAAPPP data set no. 2 is the 1976 U. S. National Survey of Young Women (John Kantner and Melvin Zelnik, John Hopkins University, principal investigators). Table 4 contains the Topic–by–Type Matrix for this data set. The matrix allows the user to see at a glance where the "areas of richness" of the data set lie. For example, we can see in the last column of Table 4 that this particular data set has a total of 386 variables; the data set is rich in information on family characteristics (142 variables), contraceptive information (56 variables), and child–bearing related information (42 variables). There are seven items relating to abortion, the first topic in the alphabetically–ordered topic list. The first row of Table 4 shows that all seven of these variables are attitudinal.

Information of the type contained in Tables 4 and 5 can be extremely helpful in ascertaining whether a given data set can be used to answer a particular research question. It is important to note that virtually no extra processing time, beyond the routine procedures used by any social scientist to create an SPSS–like set–up for his data set, is required in order to produce and display the information.

When variable names and labels from all the data sets in DAAPPP are used as input, the same program provides a matrix and variable listing that depicts the State–of–the–Archive. The 45 data sets currently in the DAAPPP collection contain 14,216 variables, characterized as shown in Table 6. While a listing of these 14,216 variables is too long to print here, such a list exists, is publicly available (for the cost of reproduction), and is updated quarterly.

Although the DAAPPP project is by no means over (the collection is currently growing at the rate of about five data sets per quarter), we see from Table 6 that there appears to be relatively little empirical data on important topics such as sexually transmitted disease and substance abuse (in the context of adolescent pregnancy studies). At the end of the DAAPPP contractual period in September 1987, we will be in a position to evaluate the amount and types of information available on adolescent pregnancy, pregnancy prevention, and family planning, and to identify significant gaps.

Social science archival data can be used in many different ways: (1) for secondary analysis (the analysis of data for purposes other than those for which the information was originally collected); (2) for meta analysis (the analysis of data common to a number of data sets to investigate similarities and differences in the patterning of relationships); (3) for longitudinal analysis of panel data (such as that found in DAAPPP Data Sets 20–24, the National

Longitudinal Survey of Youth); (4) for cross-sectional trend analysis of related surveys (such as analysis of trends in information found in DAAPPP Data Sets 11-18, the 1977, 1980, and 1982 Current Population Surveys); (5) for provision of contextual variables to add to an individual-level data file (for example, one could add all or part of the information contained in DAAPPP Data Set 8 on State Policy Determinants of Teenage Childbearing to one's individual-level data file to study the additive and interactive effects of individual versus environmental factors in producing fertility-related behavior); (6) for derivation of comparison group data against which to compare data from clinic patients or service program participants; and (7) for instructional purposes, as an exciting aid in the teaching of statistics and research design.

It is our hope that those interested in studying problems of adolescent pregnancy and family planning will use DAAPPP, and that the DAAPPP experience will be helpful in stimulating and facilitating the formation of other, special-purpose data archives containing the best scientific data on important issues facing us all.

Acknowledgements

The Data Archive on Adolescent Pregnancy and Pregnancy Prevention is funded by Contract 282-84-0083 between the Office of Population Affairs, Office of the Assistant Secretary for Health, and Sociometrics Corporation (J. J. Card). ■

Table 1

TABLE 1

LIST OF DATA SETS CURRENTLY IN DAAPPP

Data Set Id	Data Set Name (Investigators)
01	1971 U.S. National Survey of Young Women: Selected Variables (M. Zelnik & J.F. Kantner)
02	1976 U.S. National Survey of Young Women (J.F. Kantner & M. Zelnik)
03	Project TALENT: Consequences of Adolescent Childbearing for the Young Parents' Future Life, 1960-1974 (J.J. Card)
04	Detroit Mother-Daughter Communication Patterns: Mother File, 1978 (G.L. Fox)
05	Detroit Mother-Daughter Communication Patterns: Daughter File, 1978 (G.L. Fox)
06	Philadelphia Collaborative Perinatal Project: Economic, Social, and Psychological Consequences of Adolescent Childbearing, 1959-1965 (J. Marecek)
07	Nashville General Hospital Comprehensive Child Care Project, 1974-1976: Selected Variables (H>M> Sandler)
08	State Policy Determinants of Teenage Childbearing, 1979 (K.A. Moore)
09	1980 U.S. Survey of Services Provided by Adolescent Pregnancy Programs (JRB Associates)
10	1982 Evaluation of DAPP Adolescent Pregnancy Programs (M. Burt)
11	1980 U.S. Current Population Survey: Selected Variables -- Women (Bureau of the Census)
12	1980 U.S. Current Population Survey: Selected Variables -- Men (Bureau of the Census)
13	1980 U.S. Current Population Survey: Selected Variables -- Children (Bureau of the Census)
14	1982 U.S. Current Population Survey: Selected Variables -- Women (Bureau of the Census)
15	1982 U.S. Current Population Survey: Selected Variables -- Men(Bureau of the Census)
16	1982 U.S. Current Population Survey: Selected Variables -- Children Bureau of the Census)
17	1977 U.S. Current Population Survey: Selected Variables -- Women (Bureau of the Census)
18	1977 U.S. Current Population Survey: Selected Variables -- Men (Bureau of the Census)
19	First U.S. Health and Nutrition Examination Sruvey (HANES), 1971-1975 (National Center for Health Statistics)
20-24	National Longitudinal Study of Youth (NLSY), 1979-1982: Selected Variables (Waves 1-4), and Supplementary Variables (Ohio State University)
24	1981 U.S. Survey of Title X - Funded Family Planning Clinics (R. Herceg-Baron)
26	1982 National Survey of Family Growth (NSFG), Cycle III -- Women Aged 15-44 (National Center for Health Statistics)
27	1982 National Survey of Family Growth (NSFG), Cycle III -- Women Aged 15-44 (National Center for Health Statistics)
28	1979-1980 U.S. Survey of Unmarried Women Under 18 in Family Planning Clinics (A. Torres)
29	Effects of Organized Family Planning Programs on U.S. Adolescent Fertility (J.D. Forrest)
30	Johns Hopkins Study of Repeat Adolescent Pregnancy, 1976-1982 (J.B. Hardy)
31	1972-74 Ventura County of Unmarried Pregnant Women aged 13-20 (M. Eisen)
32	1982 San Jose, California Study of Adolescent Perinatal Risk Behavior (P.A. Hensleigh & N. Moss)
33	1981-1982 Evaluation of DAPP Adolescent Pregnancy Programs: Individual Level Data I (M.R. Burt)
34	1981-1982 Evaluation of DAPP Adolescent Pregnancy Programs: Individual Level Data I (M.R. Burt)
35	1979-1981 Philadelphia Study of Psychological Factors Associated With Adolescent Fertility Regulation -- Females (E.W. Flaherty & J. Marecek)
36	1979-1981 Philadelphia Study of Psychological Factors Associated With Adolescent Fertility Regulation -- Males (E.W. Flaherty & J. Marecek)
37-38	The National Survey of Children, 1976 (Child Trends, Inc.)
39	Florida-Puerto Rico Study of Adolescent Pregnancy and Neonatal Behavior, 1978 (B.M. Lester)
40	Maricopa County, Arizona Study of Child Maltreatment Risk Among Adolescent Mothers, 1976-1978 (F.G. Bolton, Jr.)
41	1955 Growth of American Families: Married Women (A. Campbell, P.K. Whelpton, & J.E. Patterson)
42	1955 Growth of American Families: Single Women (A. Campbell, P.K. Whelpton, & J.E. Patterson)
43	1960 Growth of American Families (A. Campbell, P.K. Whelpton, & J.E. Patterson)
44	1979 U.S. National Survey of Young Women (M. Zelnik & J.f. Kantner)
45	1979 U.S. National Survey of Young Men (M. Zelnik & J.f. Kantner)

Table 2

Table :

LIST OF TOPICS AND THEIR TWO-LETTER CODES

AB	Abortion	MP	Marriage patterns
AC	Agency Character-	MH	Mental health istics
AD	Adoption	ME	Meta level
AG	Age	NU	Nutrition
BF	Biological function	OC	Occupation and development
CB	Childbearing	OT	Other
CR	Childrearing	OW	Out-of-wedlock parenthood
CL	Clinical activities	PE	Personality
CM	Communication	RA	Race/ethnicity
CN	Contraception	RC	Recreation
CI	Crime	RL	Religion
ED	Education	RS	Residence/Location
FH	Family and household	SE	Sex education characteristics
FS	Friends and social	SX	Sexuality activities
GR	Gender and gender	SD	Sexually transmitted role disease
GC	Guidance and	SA	Substance abuse counseling
HL	Health	UN	Undocumented
IN	Intellectual function	WF	Wealth, finances, and material things
IV	Interview		

Table 3

LIST OF TYPES AND THEIR ONE-LETTER CODES

A	Attitudes
B	Behavior
C	Cognitions
E	Emotions
H	History
I	Intentions
M	Motivations
O	Other
P	Program/Policy
R	Reasons
S	Status
T	Traits
U	Undocumented
X	Meta
Y	Aggregate
Z	Household

Table 4

OVERVIEW OF CONTENTS, THE 1976 NATIONAL SURVEY OF YOUNG WOMEN

TOPIC	TYPE										
	ATTITUDES	BEHAVIORS	COGNITIONS	HISTORY	INTENTIONS	NOTIVATIONS	OTHER	REASONS	STATUS	META	TOTAL
ABORTION	7	0	0	0	0	0	0	0	0	0	7
ADOPTION	2	0	0	0	0	0	0	0	0	0	2
AGE	0	0	0	2	0	0	0	0	2	0	4
BIOL FUNCT	0	0	0	1	0	0	0	0	0	0	1
CHILDBEAR	1	0	2	36	0	1	0	2	0	0	42
COMMUNIC	3	0	0	0	0	0	0	0	0	0	3
CONTRACEPT	0	0	0	42	0	3	0	12	2	0	56
CHILDREAR	1	0	0	0	0	0	0	0	0	0	1
EDUCATION	1	0	0	0	3	0	0	0	3	0	7
FAMILY CHAR	2	1	1	25	2	0	0	0	111	0	142
FRIENDS	2	2	1	1	0	0	0	0	0	0	14
META	0	0	0	0	0	0	0	0	0	11	11
MARRIAGE	1	0	0	14	2	0	0	5	1	0	23
OCCUPATION	2	1	0	0	0	2	0	1	3	0	9
OTHER	0	0	0	0	1	0	1	0	0	0	2
OUT WEDLOCK	2	0	5	0	0	0	0	0	0	0	7
RACE ETH	0	0	0	0	0	3	0	0	3	0	3
RELIGION	1	1	0	0	0	0	0	0	2	0	4
RESIDENCE	1	0	0	6	0	0	0	0	0	0	15
SEX EDUC	0	0	0	6	0	0	0	0	0	0	6
SEXUALITY	3	2	0	14	0	0	0	0	0	0	19
FINANCES	1	0	0	2	0	0	0	0	5	0	8
TOTAL	30	7	9	149	8	5	1	20	146	11	386

Table 5

LIST OF VARIABLES, BY TOPIC.
NATIONAL SURVEY OF YOUNG WOMEN
PAGE 1 OF 10

TOPIC:ABORTION

NEWID	OLDID	TYPE	LABEL
ABAD2110		ATTITUDES	ABORT OK IF THE WOMAN HAD BEEN RAPED
ABAD2110		ATTITUDES	ABORT OK FOR VERY YOUNG PERSON
ABAD2112		ATTITUDES	ABORT OK IF PG ENDANG WOMANS HEALTH
ABAD2113		ATTITUDES	ABORT OK IF CHILD BORN DEFRMD OR MENTLY DEFEC
ABAD2114		ATTITUDES	ABORT OK IF THE WOMAN COULDNT AFFORD IT
ABAD2115		ATTITUDES	ABORT OK ANY REASON IMPORTANT TO HER
ABAD2116		ATTITUDES	VIEWS ABOUT HAVING AN ABORTION

TOPIC:ABORTION

NEWID	OLDID	TYPE	LABEL
ADAD2099		ATTITUDES	IF UNABLE HAVE WANTED CHIDRN, WLD ADOPT?
ADAD2100		ATTITUDES	WOULD R ADOPT CHLD INSTEAD OF HAVING OWN

TOPIC:AGE

NEWID	OLDID	TYPE	LABEL
AGH02021		HISTORY	YR OF BIRTH
AGH02022		HISTORY	MONTH OF BIRTH
AGS02003		STATUS	AGE
AGS02061		STATUS	SCREEN--AGE

TOPIC:BIOL FUNCT

NEWID	OLDID	TYPE	LABEL
BFH02132		HISTORY	AGE LIST PERIOD

TOPIC:CHILDRENBEAR

NEWID	OLDID	TYPE	LABEL
CBAD2101		ATTITUTES	IDEAL AGE FOR A GIRL TO HAVE 1ST BABY
CBAD2133		COGNITIONS	KNOW WHEN PREG IS MOST LIKELY TO OCCUR
CBAD2135		COGNITIONS	WHEN PREG IS MOST LIKELY TO OCCUR
CBHO2006		HISTORY	PREGNANCY STATUS AT MARRIAGE IND
CBHO2230		HISTORY	EVER BEEN PREGNANT
CBHO2232		HISTORY	NUMBER OF PREVIOUS PREGNANCIES
CBHO2234		HISTORY	OUTCOME OF 1ST PGAT MARRIAGE IND
CBHO2237		HISTORY	WHAT FIRST PREGNANCY THINK GOOD CHANCE
CBHO2238		HISTORY	YR OF OUTCOME 1ST PG
CBHO2239		HISTORY	MONTH OF OUTCOME 1ST PG
CBHO2240		HISTORY	AGE AT OUTCOME 1ST PG
CBHO2242		HISTORY	OUTCOME 2ND PG
CBHO2243		HISTORY	WHAT 2ND PREGNANCY THINK GOOD CHANGE

Table 6

THE STATE-OF-THE-ARCHIVE (9/85)

TOPIC	TYPE																	
	ATTITUDES	BEHAVIORS	COGNITIONS	EMOTIONS	HISTORY	INTENTIONS	MOTIVATIONS	OTHER	POLICY	REASONS	STATUS	TRAITS	UNDOCUMENTED	META	AGGREGATE	Z	TOTAL
ABORTION	24	2	56	0	79	3	0	0	8	25	0	0	0	1	16	0	217
AGENCY CHAR	0	0	0	0	40	0	0	0	247	2	24	0	0	0	0	0	213
ADOPTION	2	0	0	0	0	1	0	0	19	0	0	0	0	0	0	0	22
AGE	0	0	0	0	51	0	0	0	12	0	114	0	0	8	27	0	212
BIO FUNCT	0	174	2	3	62	2	0	0	6	3	141	0	0	12	0	0	405
CHILDBEAR	59	11	23	6	1100	70	69	0	101	78	149	0	0	7	35	0	1708
CRIME	0	8	0	0	59	0	0	0	0	0	0	0	0	0	1	0	66
CLINICAL	0	0	4	0	853	2	0	0	87	1	5	0	0	0	8	0	956
COMMUNIC	22	107	28	33	35	0	25	0	0	0	1	0	0	0	0	0	251
CONTRACEPT	51	81	549	1	769	12	6	0	58	154	3	0	0	0	15	0	1701
CHILDREAR	54	148	25	5	12	12	46	0	52	7	94	1	0	0	0	0	456
EDUCATION	19	20	2	16	355	34	18	0	136	40	178	1	0	0	12	0	732
FAMILY CHAR	38	107	43	14	419	11	25	0	95	11	1080	3	0	20	1	0	1467
FRIENDS	32	76	21	2	71	9	5	0	0	3	90	0	0	0	0	0	309
GUIDANC & COUNSE	0	0	0	0	0	0	0	0	80	0	0	0	0	0	0	0	80
GENDER	57	11	7	1	0	0	0	0	9	0	70	2	0	0	6	0	162
HEALTH	0	34	6	4	112	0	0	0	26	42	319	1	0	0	0	0	474
INTELLECT	0	14	6	0	0	0	0	0	0	0	24	129	0	1	0	0	172
INTERVIEW	2	26	12	2	15	0	0	0	0	14	6	0	0	54	0	0	129
MEA	0	0	0	0	0	0	0	0	0	0	0	0	2	681	0	0	683
MENTAL REA	0	1	2	1	18	0	0	0	4	4	14	0	0	0	0	0	44
MARRIAGE	17	0	4	1	373	17	7	0	0	23	98	0	0	3	11	0	552
MIGRATION	0	8	0	0	16	0	0	0	40	0	6	0	0	0	0	0	70
OCCUPATION	61	24	28	63	260	47	89	0	46	161	263	0	0	2	23	0	1053
OTHER	8	5	7	4	0	7	1	1	22	0	7	2	0	0	0	0	64
OUT WEDLOCK	8	1	8	0	0	0	0	0	0	0	0	0	0	0	10	0	27
PERSONALITY	14	44	0	15	3	0	5	0	0	0	0	319	0	0	0	0	400
RACE ETH	0	0	0	0	16	0	0	0	22	0	89	0	0	0	78	0	205
RECREATION	0	12	0	0	4	0	0	0	0	1	7	0	0	0	0	0	24
RELIGION	7	28	1	0	68	0	0	1	0	0	29	0	0	0	0	0	134
RESIDENCE	1	0	49	1	118	0	2	0	22	7	119	0	0	4	15	0	338
SUBST ABUSE	0	1	2	0	11	0	0	0	0	2	0	0	0	0	0	0	16
SEX TRANS DISEAS	0	0	8	0	5	0	0	0	32	0	0	0	0	0	0	0	45
SEX EDUC	0	0	0	0	26	0	0	0	14	0	0	0	0	1	1	0	42
SEXUALITY	30	29	19	14	93	8	1	0	0	2	8	0	0	0	0	0	210
UNDETERMINED	0	0	0	2	0	0	0	0	0	0	0	0	15	1	0	0	18
FINANCES	2	6	1	1	130	2	1	0	72	1	219	0	0	0	25	3	463
TOTAL	466	984	907	189	5051	317	300	2	1101	545	3052	458	17	595	247	3	14217