Using SPIRES: The ICPSR Experience

by Janet K. Vavra[1]
Inter-University Consortium for Political and Social Research
The University of Michigan

The Inter-University Consortium for Political and Social Research (ICPSR) has prepared a number of databases in the SPIRES (Stanford Public Information Retrieval System) database management system. These databases are ICPSR GUIDE, ICPSR VARIABLES, ICPSR ROLLCALLS and SMIS. Below is a description of each of these databases. There is also a brief discussion of how each of the databases was developed.

[1]Paper presented at the International Association for Social Science Information Service and Technology (IASSIST) Conference held in Marina Del Rey, California, May 21-24, 1986

The ICPSR GUIDE database contains information about the data collections in the ICPSR archive. The database contains the data holdings section of the Guide to Resources and Services, which is published annually. The quarterly additions and updates to the holdings that are announced in the ICPSR Bulletin and ICPSR "hotline" are incorporated into the database after each announcement.

The ICPSR VARIABLES database includes complete text for all questions used in selected surveys in the ICPSR holdings. Included at the time of this writing are Euro-barometers 3-21, American National Election Studies, 1948-1984, CBS News/New York Times Polls, ABC News/Washington Post Polls and selected aging and health-related studies. In addition to question texts, variable codes and frequencies, where available, are included.

The ICPSR ROLLCALLS database contains information on roll call votes taken in the last twenty years of the United States Congress. In addition to the roll call items, yeas and nays are recorded for each roll call.

The SMIS database has entries from the Survey Methodology Information System which was developed by the Bureau of the Census. Included are citations of journal publications, Bureau of the Census documents, publications, research reports, and conference papers that deal with methodological aspects of the design and conduct of surveys. There are citations in the database through 1982.

The databases are part of several on-site services available to users through the Consortium Data Network (CDNet). While giving users immediate access to information about the holdings, the databases also allow easier management of information about these holdings. Currently the archive has over 20,000 machine-readable files in approximately 1,400 study collections [these holdings represent over 6 million variables]. This number continues to

grow as nearly 150 titles are added to the holdings annually.

Without automation it is difficult to identify the collections in the holdings and to get any understanding of their contents. The ICPSR GUIDE, ICPSR VARIABLES and ICPSR ROLL CALLS databases were developed to help facilitate access to more complete and reliable information about the holdings. Further, they seek to provide this access in a timely and economical manner.

An appendix to this paper lists the indexes and elements found in each of these databases. All the databases can be accessed on-line through CDNet or can be installed locally either in SPIRES or converted into other database management systems.

Database preparation

All the input information used in each of the four ICPSR databases was already machine-readable so there was no need to automate large quantities of raw information. However, as one would expect, none of the machine-readable files were in a form that could be input directly into the respective SPIRES databases. The input files for each database had to be reformatted to make them acceptable to SPIRES.

The ICPSR Guide to Resources and Services has been a machine-readable text file since about 1973. This file is updated and used each year to prepare the annual hardcopy document. It was this file that was used as input for the ICPSR GUIDE database. The file was prepared for SPIRES input with a series of editing procedures using the Edit capability on the ICPSR PRIME minicomputer. The edit procedures removed characters from the text that are unacceptable to SPIRES, such as semi-colons (SPIRES uses semi-colons as terminators), and replaced them with acceptable characters. Using the print control characters that existed in the original file as a guide, the edit procedures inserted the appropriate SPIRES element names and terminated the fields with semi-colons. These procedures handled most of the conversion into SPIRES format. The SPIRES programs that enter the information into the database identified any remaining problems. These were handled on-line as the entries were added to the database.

Information in the ICPSR VARIABLES and ICPSR ROLLCALLS databases comes from OSIRIS machine-readable codebooks. Before these files are acceptable to SPIRES, they must go through several steps. The first step is a simple set of editing procedures that search for unacceptable characters and replace them with acceptable ones. The second step creates several files which handle such conditions as references to question numbers, for which full question text is supplied earlier in the codebook, and references to notes found at the back of the codebook. This step results in three files: the question text file, a note file and the codebook file now without the notes at the end. The third and final step performs several operations:

1. it merges the appropriate question text into the variable description where references to the question text are made, and when it is possible to locate the question number in the codebook,

2. it inserts appropriate note information where references to given notes appear and

3. it generates the appropriate element names and delimiters to make the file acceptable to SPIRES. Final error checking is done by the SPIRES input programs as the reformatted codebook files are input into the database.

The original information for the SMIS database was supplied on several magnetic tapes by the Bureau of the Census. The information on each tape was written onto a line file and editing procedures were set up to make the information compatible with SPIRES. Each file was scanned for characters unacceptable to SPIRES. These were then changed to acceptable ones as they were encountered. In order to facilitate the conversion, ICPSR retained the element names assigned by the Bureau of the Census and also followed the database structure for SMIS that had been developed by the Bureau of the Census. The edited files were used as input into the SPIRES database.

Currently the ICPSR GUIDE database contains approximately 1,400 entries, the ICPSR VARIABLES database has over 47,000 entries, the ICPSR ROLLCALLS database has approximately 18,000 entries and the SMIS database contains over 7,000 entries. All of the databases remain dynamic as more records are added and as some existing ones are updated.

Database features

In order to facilitate the searching of the databases, SPIRES was instructed to create "indexes" for each database. The indexes consist of individual words that are found in a particular element or set of elements. The indexes are generated by SPIRES as entries are added to a database. It should be noted that this type of index construction is not like a manually-generated classification scheme. Since the indexes are "word" indexes, search terms (or values) cannot be longer than one word. Two or more search commands, each using a word, are used when searching for a phrase.

The databases accept truncated value searches, allowing the user to obtain a valid search result without complete information for a given item. A user may not know an author's full name, but by inserting a "#" sign at the end of the search command line, the user can instruct SPIRES to save all records that begin with the value listed in the search command. For example, if the user is not certain whether an author's name is Williams or Williamson, by asking SPIRES to retrieve all authors with the name Williams#, the user will get all authors whose last name begins with Williams including Williamson.

The databases accept search commands in either upper or lower case or in mixed mode. This frees the user from having to remember the appropriate mode for the database or running the risk of invalid search results simply because commands were given in the wrong case.

Selected special characters in the body of the text are ignored as a user searches the databases. This allows the user to locate all entries, for example, with California in the title, irrespective of whether California is listed alone or in brackets or parentheses.

Each database has custom formats that display the results of searches in a more readable form than the routine SPIRES default format. Output can further be processed through these formats so that it is more readable by sorting a string of values, by forcing output from a given field to upper or lower case, by providing more meaningful headings for fields, or by rearranging the order in which fields are output so that the information fits together in a more logical manner.

Users who are not happy with the indexes that have been created can ignore index searching and switch to sequential searching. Since in sequential searches all information in each entry is examined in sequence to determine whether or not it meets the search criteria, this method of searching is more time consuming and expensive. Frequently users switch to sequential

searching as part of an index search rather than in place of the index search. Being able to use both modes of searching gives the user more flexibility and also different ways to test the results of searches.

Conclusion

The ICPSR experience with the SPIRES database management system has been favorable. The capacity of the system is basically unlimited so that databases can continue to expand without much fear that the maximum number of entries will be exceeded. In fact, it is more likely that the capacities of the storage devices will sooner be exhausted. The system has many capabilities and yet can be used quickly and effectively by inexperienced users. Finally, it has not been an especially difficult or expensive task to prepare existing machine-readable information into input files that are acceptable to SPIRES.□

appendix

ICPSR SPIRES Databases

Below is a list of the four ICPSR SPIRES databases that are available through CDNet. A brief description of each database, a list of the indexes which can be used to search each one and a list of the elements found in each is included.

ICPSR GUIDE Database

CONTENTS: Archival holdings section of the Guide to Resources and Services issued and collections announced in Bulletins that have been issued since the publication of the Guide. Database is most up-to-date list of ICPSR data holdings.

INDEXES:

```
Goal Records:  COLLECTION
Simple Index:  AD, ADDED
Simple Index:  UP, UPDATED
Simple Index:  T, TITLEW, TITLEWORD, TW, TWORD
Simple Index:  S, SUBJECTWORD, SUBWORD, SW, SWORD
Simple Index:  P, PIAUTH, PINAME, PIW, PIWORD, PW
```

ELEMENTS:

```
Key     :  STUDYNO, ICPSRNO, IDNO, MRDF.NUM, STUDNO, 037A
Element:   DATE-ADDED, DA, DAD, DADD, DATEADD
Element:   DATE-UPDATED, DATEUP, DU, DUP, DUPD
Element:   INVESTIGATOR, INVEST, PI, PRIN.INVEST, 199A
Element:   TITLE, TI, TITL, 245A
Element:   SUMMARY, DESC, DESCRIP, SUM, 520B
Element:   SUBJECT.TERM, S.TERM, STERM, SUB, SUB.TERM, SUBJ, SUBJ.TERM,
             SUBJECT, 653A
Element:   ARCH.FILTER, A.FILTER, AFILTER, ARCH, ARCHFILTER, ARCHIVE,
             FILTER

Structure:  CLASSIF
  Element:  ICPSR.CLASSIF1, CLASSIF1, 972A
  Element:  ICPSR.CLASSIF2, CLASSIF2, 972B
  Element:  ICPSR.CLASSIF3, CLASSIF3, 972C
  Element:  ICPSR.CLASSIF4, CLASSIF4, 972D
  Element:  ICPSR.CLASSIF5, CLASSIF5, 972E

Element:   EXTENT.COLLECT, E.COL, E.COLLECT, ECOL, ECOLLECT, EXT, EXTENT,
             EXTENT.FILE, EXTENTCOLLECT, FILES, 300A
Element:   SERIES.NAME, S.NAME, SER, SER.NAME, SERIES, SERIESNAME, SNAME,
             440A
Element:   SERIES.INFO, S.INFO, SER.INFO, SERIESINFO, SINFO, 500A
Element:   RESTRICTIONS, LIMITATIONS, LIMITS, REST, RESTRICT, 506A
Element:   DATA.TYPE, D.TYPE, DATATYPE, DTYPE, 516A
Element:   TIME.PERIOD, CHRON.COVER, T.PERIOD, TIMEPERIOD, TPER, TPERIOD,
             523A
Element:   DATE.OF.COLLECT, COLLECT.DATE, D.COLL, D.COLLECT, DATADATES,
             DATESOFCOLLECT, DCOLL, DCOLLECT, 523B
```

appendix

```
Element:  FUNDING.AGENCY, F.AGENCY, FAGENCY, FUND, FUNDING,
            FUNDINGAGENCY, SPONSOR, 536A
Element:  GRANT.NUMBER, G.NUMBER, GNO, GNUMBER, GRANT, GRANT.NUM,
            GRANTNO, GRANTNUMBER
Element:  DATA.SOURCE, D.SOURCE, DATASOURCE, DSOURCE, SOURCE.DATA, 537A
Element:  DATA.FORMAT, D.FORM, D.FORMAT, DFORM, DFORMAT, FORM, FORMAT,
            538A
Element:  COLLECT.NOTE, C.N, C.NOTE, CNOTE, COLL.NOTE, COLLECTNOTE,
            COLLNOTE
Element:  SAMPLING, SAM, SAMP, 567A
Element:  UNIVERSE, UNI, UNIV, 567B
Element:  RELATED.PUBS, PRIMARY.PUB, PRIMARY.PUBS, R.PUBS, REL.PUB,
            REL.PUBS, RPUBS, 581A
Element:  CLASSNO, CLASS, CLASSNUM, CNO, CNUM, ICPSR.CLASS, 962A

Structure:  PART
  Element:  PARTNO, PNO, 245P
  Element:  PART.NAME, P.NAME, PARTNAME, PNAME, 565P
  Element:  FILE.STRUCT, F.STRUCT, FILESTRUCT, FSTRUC, FSTRUCT, STRUC,
              STRUCT, 538S
  Element:  CASE.COUNT, C.COUNT, CASECOUNT, CASES, CCOUNT, 565A
  Element:  VARIABLE.COUNT, V.COUNT, VAR, VARIABLECOUNT, VARIABLES, VARS,
              VCOUNT, 565B
  Element:  LRECL
  Element:  RECORDS.PER.CASE, R.CASE, R.PER.C, RCASE, REC.PER.CASE,
              RECORDS.CASE, RECORDSPERCASE, RECPERCASE, RECS.PER.CASE,
              RECSPERCASE, RPC, RPERC, 565C
```

appendix

ICPSR VARIABLES Database

CONTENTS: Descriptions and codes of variables found in the American National Election Studies 1948-1984, Euro-barometers 3-21, Quality of American Life, 1971 and 1978, media polls, and health- and aging-related surveys.... Database under development as studies continue to be added.

INDEXES:

```
Goal Records:  VARIABLE
Simple Index:  ICPSRNUM, ID, STUDYNUM
Simple Index:  AD, ADDED
Simple Index:  UP, UPDATED
Simple Index:  CHRONOLOGICAL, COVERAGE, PERIOD, TIME, TIMEPERIOD
Simple Index:  S, SUB, SUBJECT, V, VAR, VARIABLE
```

ELEMENTS:

```
Key     :  ID-NUM
Element:  DATE-ADDED, DA, DADD, DATEADD
Element:  DATE-UPDATED, DATEUP, DU, DUP
Element:  STUDYNUM, SN, SNUM
Element:  STUDYDSNUM, DATASET, DS
Element:  VARNUM, V#, VNUM
Element:  VARNAME, VN, VNAME
Element:  TIME-PERIOD, TIME, YEARS
Element:  VARDESC, DESC, DESCRIPTION, Q, QUESTION
Element:  VARCODES, CODE, CODES
Element:  VARFREQS, FREQ, FREQS
Element:  STUDY-NAME
```

appendix

ICPSR ROLLCALLS Database

CONTENTS: Variables found in the machine-readable collection of the United States Congressional Roll Call Voting Records [ICPSR 0004]. Contains most recent Congressional Sessions. Database under development as Congresses continue to be added.

INDEXES:

```
Goal Records:  ROLL.CALL
Simple Index:  DS, ICPSRDS, ID
Simple Index:  AD, ADDED
Simple Index:  UP, UPDATED
Simple Index:  CHRONOLOGICAL, COVERAGE, PERIOD, TIME, TIMEPERIOD
Simple Index:  S, SUB, SUBJECT
Simple Index:  DATE.VOTED, V, VOTED
```

ELEMENTS:

```
Key     :  ID-NUM
Element:  DATE-ADDED, DA, DADD, DATEADD
Element:  DATE-UPDATED, DATEUP, DU, DUP
Element:  STUDYNUM, SN, SNUM
Element:  STUDYDSNUM, DATASET, DS
Element:  DSDATES, DSD
Element:  VARNUM, V#, VNUM
Element:  VARNAME, VN, VNAME
Element:  RCSOURCE, RCS, SOU
Element:  RCDATE, RCD, RCDAT
Element:  RCNUM, RCN
Element:  RCPROPOSER, RCP, RCPROP
Element:  RCVOTES, RCV, VOTES
Element:  VARDESC, DESC, DESCRIPTION, M, MOTION
Element:  RCNOTENUM
Element:  NOTE-TEXT, NN, NOTE
```

appendix

SMIS Database

CONTENTS: Contains entries from the SMIS (Survey Methodology Information System) that was originally developed by the Bureau of the Census. Included are journal publications, Bureau of the Census publications and documents, research reports, etc. that deal with methodological aspects of the design and conduct of surveys.

INDEXES:

```
Goal Records:  TITLE
Simple Index:  AD, ADDED
Simple Index:  UP, UPDATED
Simple Index:  T, TITLEWORD, TL, TW
Simple Index:  S, SUB, SUBJECTWORD, SUBWORD
Simple Index:  A, AUTH, AUTHOR
Simple Index:  C, CON, CONFER, CONFERENCE
```

ELEMENTS:

```
Key     :  NO
Element:   DATE-ADDED
Element:   DATE-UPDATED
Element:   T-TITLE, T
Element:   DT-DOCTYPE, DT
Element:   KW-KEYWORD, KW
Element:   D-DATE, D
Element:   CJ-LITCODE, CJ
Element:   NR-NO.OF.REFS, NR
Element:   V-VOLUME, V
Element:   I-ISSUE, I
Element:   PI-PAGES, PI
Element:   FG-FINANC.SOURCE, FG
Element:   FC-AGEN.CNTRL.NO, FC
Element:   FP-PROJECT.NO, FP
Element:   PT-PART.NUMBER, PT
Element:   LN-LANGUAGE, LN
Element:   SE-SERIES.NAME, SE
Element:   DF-DOCUMENT.FORM, DF
Element:   AV-AVAILABILITY, AV
Element:   DL-DISTRIB.LIMIT, DL
Element:   RS-AGENCY.RPT.NO, RS
Element:   RM-MONITOR.NO, RM
Element:   RO-DISTRI.RPT.NO, RO
Element:   AB-ABSTRACT, AB
Element:   CN-CONFERENCE, CN
Element:   PU-PUBLISHER, PU
Element:   VR

Structure:  AU
  Element:  NA-AUTHOR, NA
  Element:  RA-EDITOR, RA
  Element:  ON-AUTHAFFIL, ON
```

appendix

```
Structure:  PARENT
  Element:  MT-SOURC.TITLE, MT
  Element:  MD-SOURC.DATE, MD
  Element:  MKW-SOURC.KEYWRD, MKW
  Element:  MNR-SOURC.NO.REF, MNR
  Element:  MV-SOURC.VOLUME, MV
  Element:  MI-SOURC.ISSUE, MI
  Element:  MPI-SOURC.PAGES, MPI
  Element:  MFG-SOURC.SUPPOR, MFG
  Element:  MSE-SOURC.SERIES, MSE
  Element:  MDF-SOURC.DOCFRM, MDF
  Element:  MDL-DISTRIB.LIM, MDL
  Element:  MAV-SOURC.AVAIL, MAV
  Element:  MRS-SOURC.AGENNO, MRS
  Element:  MRM-SOURC.MON, MRM
  Element:  MRO-SOURC.DISTRI, MRO
  Element:  MAB-SOURC.ABSTR, MAB
  Element:  MCN-SOURC.CONFER, MCN
  Element:  MPU-SOURC.PUB, MPU
  Element:  MPT-SOURC.PARTNO, MPT
  Element:  MLN-SOURC.LANG, MLN
  Element:  MFC-SOURC.CTRLNO, MFC

Structure:  MAU
  Element:  MNA-SOURC.AUTHOR, MNA
  Element:  MRA-SOURC.EDITOR, MRA
  Element:  MON-SOURC.AFFILI, MON
```