A Review of OSIRIS IV[1]

Richard C. Roistacher
Bureau of Social Science Research
Washington, DC 20036

April 1, 1980

The Survey Research Center at the University of Michigan's Institute for Social Research has begun the dissemination of OSIRIS IV, the latest version in Michigan's series of statistical systems. OSIRIS IV is designed to be an eventual replacement for the OSIRIS III system. OSIRIS has long been the ugly duckling of statistical systems. Data analysts find its six volumes of documentation forbidding and unintelligible, and dislike the cross-tabulation program's control syntax and lack of category labels. As a result, most data managers have remained unaware of OSIRIS III's unequalled power at managing sequential data files.

Documentation. The OSIRIS IV statistical system is documented by "OSIRIS IV Statistical Analysis and Data Management Software System," a loose-leaf manual of approximately 200 pages, written by the staff of the Survey Research Center Computer Support Group. Additional documentation for some of the structured file facilities is contained in a 138-page book, "OHDS: Introduction to OSIRIS Hierarchical Data Structures" produced by the computer support group of the center for political studies. The official OSIRIS IV manual is a reference work, much in the tradition of the OSIRIS III documentation, while the OHDS volume takes a more tutorial approach.

To the user, OSIRIS IV appears much like its predecessor. All programs (which have been relabelled "commands") run under a monitor, which handles control language and file assignments. The program setup file, while no longer restricted to 80 characters as in OSIRIS III, is still oriented to line by line input.

Procedures. As of the moment of this writing, OSIRIS IV consists of 32 procedures. Sixteen of the procedures are for data analysis, ten for general data management, and six for the management of structured files. The data analysis procedures include the usual routines for crosstabulation, correlation,

---

1. This work was supported by contract 78-SS-AX-0028 from the Bureau of Justice Statistics, U. S. Department of Justice.

regression, and analysis of variance. OSIRIS IV also contains SEARCH, a new version of AID and related programs. Among the more unusual offerings are PSALMS, a sampling error analyzer, and REPERR, a repeated samples program for (among other things) Tukey's Jacknife. Several nonmetric scaling routines will shortly be added to the system's repertory.

OSIRIS IV contains a full set of the data management procedures which made its predecessor so useful to data managers. Along with both preprocessor and stand-alone versions of the recode (the most powerful of any statistical system's), there are procedures for defining, copying, sorting, listing, and correcting rectangular files. A set of six programs provide facilities for defining and manipulating the structure of multi-record files. The SBUILD procedure is used to combine several rectangular files into a single structured file. MERGE and UPDATE are used to combine (or split) structured files, while STRANS is used to alter the structure of a multi-record file. There is no clear separation of functions among this set of procedures, since they were written at different times in the system's development, and by people with differing ideas about the handling of structured files.

Although an OSIRIS IV setup looks superficially like an OSIRIS III setup, there are important differences in the organization of the two systems. The design of OSIRIS IV takes into account that data transformations and structures are often reused several times over the course of a run. The system is novel in having its recodes and structure definitions defined in one place and invoked in another. (One of the more common errors in the use of OSIRIS IV is to define a RECODE and then fail to invoke it.) Once the idea of separating a procedure's definition from its invocation has been recognized, the power of the idea becomes apparent. Among the entities which are separately defined and invoked are filters, RECODES, and data structure definitions.

File building. OSIRIS IV has perhaps the best designed file-building procedure of any available system. The OSIRIS IV file-building procedure consists of two steps. The user constructs a dictionary describing the raw data, much as is done in OSIRIS III or in the SPSS DATA LIST procedure. The user can then treat the raw data file as if it were an OSIRIS internal-format file described by the dictionary. If it is desired to create a "real" OSIRIS file in which blanks have been recoded to missing data, unused variables squeezed out, etc., the dictionary and the raw data file are passed through the data transformation program. Thus, the OSIRIS IV user has the option of processing from the raw data, as well as the ability to generate a clean internal-format file.

Structured files. By far, the most important advance in OSIRIS IV, and the one which merits the immediate attention of data analysts is its ability to generate and handle structured

files. OSIRIS IV, like its predecessor, is basically a tape-oriented system, and handles its files sequentially. It is able to process hierarchical and panel study files directly, without the need for extract or work files. Although the OSIRIS documentation makes use of the words "schema" and "subschema" from the data base management literature, OSIRIS IV is not a data base management system. In a DBMS, the user manipulates the data in a structured form. The DBMS user is given a schema, a logical structure which describes a relation among records in the data base.

For example, a schema might show the records in a data base as consisting of families in which records of children are subordinate to records of parents. Another schema might show the data from the same data base as schools in which children are subordinate to classes and teachers, who in turn are subordinate to schools and principals. The user of a DBMS need not be concerned with the physical structure of the data. The schema will cause the data to be presented in a logical structure which is independent of the data's physical structure.

Structure definitions. Traditional statistical systems perform analyses on vectors of variables, and so does OSIRIS IV. The logical unit in an OSIRIS IV structured file is an "occurrence," a group of related records which describe the highest level unit of analysis. The OSIRIS hierarchical data structuring facility reduces an occurrence to an "entry," a data vector equivalent to a record in a rectangular file. Analysis programs operate only on entries, never on occurrences.

Consider the example in Figure 1. The five records in the figure constitute what is called an "occurrence," in OSIRIS IV.

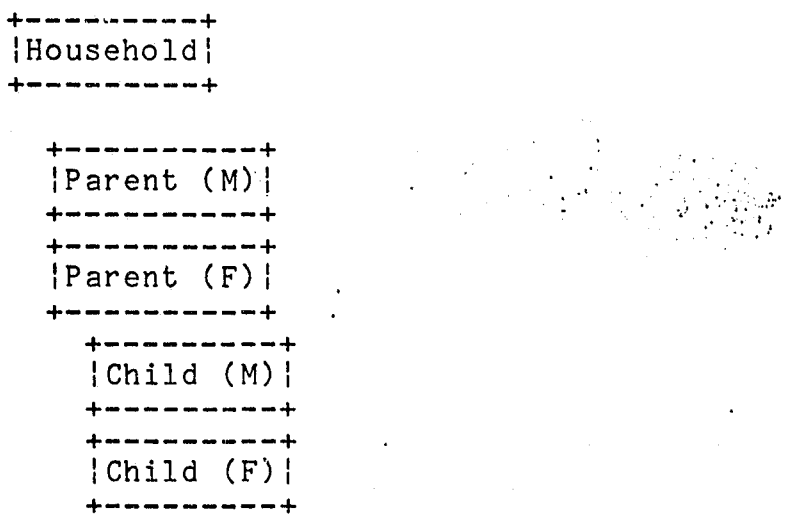

Figure 1: A Three-level Occurrence.

These records contain data about the parents, children, and house composing a household. The OSIRIS IV structure definition

facility allows this occurrence to be presented to an analysis program as any of several types of "entry." An entry is a vector equivalent to an "observation" or record in a rectangular file. Figure 2 shows some of the entries which OSIRIS IV can create from this occurrence.

A.

Household	Parent (M)	Parent (F)

B.

Household	Parent (M)
Household	Parent (F)
Household	Child (M)
Household	Child (F)

C.

Parent (M)	Child (M)
Parent (F)	Child (F)

Figure 2: Three examples of entries.

Entry A consisting of records 1, 2 and 3 defines a unit of observation which could be called "a pair of spouses in a house." In B, the same occurrence has been used to generate four entries representing "people and the houses they live in." In example C, the occurrence has been decomposed into entries of "parents with one child of the same sex."

This last example, while a little unusual, represents an important feature of the OSIRIS hierarchical data facilities; i.e., that they go far beyond the normal manipulations of bringing the higher level records down to the lower or of stringing the lower level records out at the end of the higher. OSIRIS IV has made possible some major advances in data structuring, and should have an enormous impact on the conduct of research. OSIRIS IV should make it possible for investigators to attempt multiple analyses of structured files and to approach the studies of panels with more efficiency and less trepidation.

The reader should be warned that OSIRIS IV can be a veritable bed of quicksand for its users. Like all new systems, it contains bugs, which will be discovered only through their

bite. Even when such bugs are corrected, however, the structured file facility is intrinsically more difficult to use than any statistical system which does not handle structured files.

The decomposition of the multi-record occurrence into the single record entry is accomplished through an intricate manipulation of sort keys and variable numbers.

An example. A "simple" example will show what is required to produce an entry. Consider an occurrence of the type shown above. Figure 3 shows the variables in the three record groups which make up an occurrence.

Group	Level	Group Name	V1	V2
1	1	Household	State	Value of House
2	2	Parent	Sex	Years education
3	3	Child	Sex	Verbal SAT score

Figure 3: Variables in the occurrence.

Assume that we wish to determine the relation between parents' mean level of education and children's verbal SAT scores. To do so will require the generation of an "entry," a pseudo-record, for each child, which includes the education for each parent and the child's SAT score. The program fragment in Figure 4 will produce from the occurrence an entry for each child which contains the following variables:

V2001:	Parent's sex (First parent)
V2002:	Years of education (First parent)
V2101:	Parent's sex (Second parent)
V2102:	Years of education (Second parent)
V3001:	Child's sex
V3002:	Child's SAT score

The program fragment shown in Figure 5 shows a set of procedures for regressing a child's SAT score against the years of education for the same-sex parent and years of education for the other-sex parent.

```
 1 &ENTRY
 2 ENTRY=1 UNIT=3
 3 G2+G3
 4 GNUM=2 MAXOCC=2 RENUMBER=2001 VINCR=100
 5 GNUM=3 RENUMBER=3001
 6 &RECODE
 7  RECODE=1
 8  NAME R1'YRS ED.  SAME-SEX PARENT', R2'YRS ED OPP-SEX PARENT'
 9  IF MDATA(V2001,V2002,V2101,V2102,V3001,V3002) THEN REJECT
10  IF V3001 EQ V2002 THEN R1=V2002 AND R2=V2102 -
11    ELSE R1=V2102 AND R2=V2002
12 &REGRESSN DICTIN=EDDICT DATAIN=EDDATA
13 REGRESSION OF CHILD'S SAT ON PARENTS EDUCATION
14 RECODE=1 ENTRY=1 PRINT=(CORREL,USTATS)
15 VARS=R1,R2 DEPV=V3002
16 &END
```

Figure 4: An OSIRIS IV program.

Line 1 of Figure 4 invokes the structure definition procedure.

Line 2 numbers the structure definition ("entry") and indicates that an entry is to be generated for each third level record.

Line 3 indicates that an entry is to be generated only if both a Group 2 (parent) and Group 3 (child) record have been read since the previous entry was generated.

Line 4 says that up to two parent records are to be handled, that variables in parent records are to be renumbered 1001,1002,..., and that successive occurrences of parent records will have 100 added to their variable numbers.

Line 5 renumbers variables in child records beginning with 3001. This entry definition will produce the variable numbers in the entry shown above.

The recode which is defined in lines 6-11 will reject the entry if it contains any missing data. If the entry is accepted, the recode assigns the years of education of the same-sex parent to temporary variable R1, and the years of education of the other-sex parent to R2. The regression procedure beginning on line 11 invokes the RECODE and ENTRY procedures to perform the analysis on data in a file called EDDATA, whose dictionary is in a file called EDDICT.

This example is about par for the OSIRIS IV course. It is a little complicated in that it simultaneously rectangularizes across (aggregating parents' sex and education) and down (from parents to children). This five finger exercise, however, is a model of simplicity compared to what happens when panel studies,

multiple hierarchies, and recodes are all put together in a veritable tocatta and fugue for computer. I expect that feed back from users will provide the designers with ways of cleaning up the present data structuring language. However, I fear that a true data base language will have to wait for the advent of OSIRIS V or one of its competitors. Users who wish to analyze structured files are just going to have to study hard and work through the problems.

Implications. The advent of OSIRIS IV and other systems for handling structured files has important implications for the way in which social research is done. We have evolved from a period in which social science data processing was the province of a few high priests, to a point at which most graduate students and recent PhD's feel reasonably comfortable with the computer. However, much of the conduct of research involving analysis of machine readable data files seems to consist of a senior investigator waiting with increasing frustration while a graduate student learns the fine art of tape slinging. In a successful project, one month of tape slinging generally results in one productive week of analysis. Things are often speeded up where universities and organizations have local data archives or computer support centers, that do the data management work for researchers. Technicians or highly qualified graduate students at the support center can do the data management, leaving the investigator to set up his or her own analysis runs.

The advent of structured file capabilities will lead to an ever larger role for the skilled social science computernik and for the research support center. Many researchers will be unable to write their own data structure definitions, and most will be totally unable to cope with the recoding and generation of a structured file. More than ever, researchers dealing with structured files will become clients of the technical people who can generate and manage such files.

I suspect that we will see a development much like that at the earlier stages of the social science computing, where technology starts to make things easier and where the requisite computing skills migrate from the province of a small group of technically specialized people to the larger community of their clients.

In this review, I have tried to stress both the power and the difficulty of OSIRIS IV. OSIRIS IV should be part of the library of every computing center with a data analytic clientele. Social researchers should make themselves aware of OSIRIS IV's capabilities and evaluate how those capabilities might best be employed. However, they should be warned that the path of the analyst of structured files is a steep and rocky one. Wear heavy shoes and carry a stick. (A flask of brandy is sometimes helpful.)

END