AN OVERVIEW OF OSIRIS III

by

Gregory A. Marks
Center for Political Studies
Institute for Social Research
The University of Michigan
Ann Arbor, Michigan 48109

OSIRIS III has been developed at the Institute for Social Research (ISR) to meet the needs of the research and data archive staffs, encompassing both contemporary survey research, analysis of current and historical aggregate data and active redissemination of data to a world-wide community of social scientists. OSIRIS may be used effectively for student projects and straightforward research tasks, yet the greatest virtues of the system are in the capabilities provided for the original preparation, documentation, and manipulation of relatively large and complex data files, in an assemblage of advanced analysis routines, and in the system's openness and flexibility for meeting users' needs. As the generality of the design of OSIRIS has become recognized, use has spread beyond the social sciences into such diverse realms as the management of medical records or the analysis of automobile accident data. There are over 200 installations of OSIRIS III at present, of which about half are academic sites, the balance being government or commercial installations.

OSIRIS Standard Data Files

OSIRIS was developed in an environment where a typical user has a private data file which contains hundreds or thousands of variables, and thousands or tens of thousands of cases. As a consequence, OSIRIS was designed to utilize a dictionary containing format information for each variable, so that the user does not have to re-enter this information for each computer run. This substantially reduces the staff time and boredom of run preparation, and also reduces the number of costly setup errors.

The standard OSIRIS data set consists of two files, the dictionary and the data. The data are stored as a sequential file on tape or disk, case by case. Within the records for the cases are fixed fields for the values of the variables. The format for each variable is described by an associated record in the dictionary. Some of the information in this dictionary record includes:

- an identifying number by which the user and the program reference the variable.
- a 24 character alphabetic label for the variable, used in all output for readable, redundant identification of the variable.
- format information such as field position and width, number of decimal places, whether pure numeric or alphabetic, whether character or binary mode, and so forth.
- identification of code values which signify missing data.
- on additional (optional) reccords, the user may provide "codebook" text to describe how the data for the variable were obtained or measured, the details of the coding scheme or the measurement units, and other forms of data documentation.

Figure 1 below illustrates the relationship between the dictionary and the data, for further clarity.

In addition to the dictionary/data files just described, there are also other standard file formats for matrices such as arrays of correlation coefficients, means, standard deviations, and bivariate

Figure 1

OSIRIS Dictionary and Data Illustration

	Number identifying variable	Name for variable	Format for variable	Codes for missing data
Dictionary	1	Respondent number	Position 1-4,...	(none)
	2	Age	" 5-6,...	0, ≥ 95
	3	Income	" 7-11,...	≤ -1

	1	2	3	4	5	6	7	8	9	10	11	<=== Positions
Data				1	3	5	1	4	9	5	5	<===== case or respondent 1
				2	2	1		3	7	0	0	<===== " " " 2
				3	7	2		9	7	2	0	<===== " " " 3
				4	9	9				-	2	<===== " " " 4
	Variable 1 (V1)				Variable 2 (V2)		Variable 3 (V3)					

frequency distributions. Thus output from a variety of statistical computations can be used as input to other appropriate statistical routines in OSIRIS.

To access data in a standard OSIRIS file, the user need specify only the desired variable numbers. The software utilizes the dictionary for other necessary information. Selection by the user of a subset of the cases is also possible, through what is termed a filter. Any of the variables in the data file may be combined in a logical expression to test whether or not to pass specific data cases to the OSIRIS program. An example for the file shown in Figure 1 might be "INCLUDE V2=21-99 AND V3=2000-9999," which would pass respondents 2 and 3. The selection of specific variables and/or cases can be done on a temporary basis while data are input to a specific program run, or these subsets can be made into a separate, permanent file by means of any of several programs. In such a process the dictionary is automatically properly subsetted along with any associated codebook text records. Some similar capabilities exist for standard matrix files.

Data Management

The fact that projects within ISR collect, analyze, and rediffuse to others many large bodies of data is clearly reflected in the design of OSIRIS. Thus for those situations where an archive needs to recover an ancient collection of data for reanalysis by today's scholars, a series of programs for handling multiply punched data are found in the package. Because much of the data collected at ISR starts in card form (or card-image on tape or disk), and because this is also the most common form in which data are sent to the archives of the Inter-university Consortium for Political and Social Research (located within ISR), OSIRIS has capabilities for working with card data. All OSIRIS programs that accept a standard dictionary/data file can also input a card-image file using a form of dictionary to provide format information. OSIRIS also incorporates programs which perform functions specialized to card files such as checking and correcting the match-merge of card decks, and obtaining frequencies for special non-numeric data codes. Most of these capabilities are intended as helping the user check the basic condition of old, unfamiliar, or inexplicably garbled data files and unscramble the problems confronted.

Users with some experience in data collection now know how to avoid the more basic errors of card handling and hence rarely need the programs just described. These users move directly into the generation of the standard OSIRIS dictionary and data file, a one-time process known as "file building." There is a program in OSIRIS for this specific purpose which incorporates a number of editing and sequence checks. With the OSIRIS file in place, the user has available programs which enable more sophisticated error checking, such as for wild codes and logial inconsistencies, and correction capabilities.

Getting data into proper organization for a research project often means reordering the file, adding new cases to the file, or inserting new variables for the existing cases, such as when doing repeated measurements over a time span of months or years. Thus one valuable data management function in OSIRIS is a sort/merge program which calls upon the system utility after processing a simplified user setup. Another form of merge is possible in OSIRIS, and warrants some additional description. In Figure 2 are shown two files, A containing data on parents, and B containing data on children. Dictionaries have been left out of the figure for simplicity. The function provided is the combination or merging of variables for cases which are identified as matching. In Example 1, two procedures available during the merge are illustrated: first, when a case occurs in one file but not in the other, the valid data are retained and the missing information is signified by padding with missing data codes. Secondly, whenever more than one possible match is found (i.e., child 2 of parent 1), each additional case is produced as another record along with a duplicate entry of the matching record from the other file (i.e., parent 1 appears in each output record). In Example 2, both these options are reversed; thus whenever records from both files are not found, the entire case is dropped, and all matches beyond the first to a case are dropped.

To facilitate, for example, the generation of summaries across all the members of a family or a work group, or the addition of values for small geographic areas to obtain totals for larger areas, there are several programs in OSIRIS which generate sums, means, and perform other operations across groups of records. Then if desired, one may use the previously described merge program to combine the group scores with the data in each component case

Figure 2

Merge of Variables Illustration

File A

parent 1
parent 2
parent 3

File B

child 1 of parent 1
child 2 of parent 1
child 1 of parent 3
child 1 of parent 4

Example 1. the result file C

parent 1	child 1 of parent 1
parent 1	child 2 of parent 1
parent 2	(missing data)
parent 3	child 1 of parent 3
(missing data)	child 1 of parent 4

Example 2.

parent 1	child 1 of parent 1
parent 3	child 1 of parent 3

and examine differences, case by case.

There are in OSIRIS an extensive collection of arithmetical and logical operators for performing transformations on variables, including a variety of methods for univariate or multivariate recodes. Conditional expressions for control of the sequence of operator execution are included. Complete capabilities are provided for handling of missing data during computations. The results of the operations may be output as a standard OSIRIS data file, in which case the program generates the necessary new dictionary. Alternatively, most of the transformation capabilities are available for temporary operation on the data while being input to a specific analysis program.

One other aspect of data handling in OSIRIS deserves special mention: the ability to move data from a standard OSIRIS file into another software package. Interfaces for reading OSIRIS files as such exist in SAS, SPSS, P-STAT, and other software. Where such interfaces do not exist, another alternative is to create the appropriate format statements for reading the data file, which is typically in character mode. Finally, there is a program in OSIRIS which generates card-image files from an OSIRIS dataset with minimal user setup. This program also produces a custom description of the variables in that card-image file, including a tailored codebook if such text is present in the dictionary.

There are still other, somewhat more specialized data management and manipulation functions available in OSIRIS. In combination, all the capabilities available in OSIRIS make it as much a data management system as it is a statistical package.

Statistical Capabilities

The collection of statistical routines in OSIRIS is extensive, although of course not all-encompassing. A brief listing of the major programs and functions is indicative of the breadth of capabilities; persons interested in more detail should examine the documentation listed later.

Program	Function
TABLES	Univariate and bivariate tables and measures
SCAT	Scatterplots
NTILE	N-tiles
MDC	Pearsonian correlations
PARTIALS	Partial correlations
CORREL	Ordinal and categorical correlations
REGR	Linear regression with 2-stage
FMEANS	Variance analysis, one-way
MANOVA	Variance analysis, multivariate
AID3	Automatic Interaction Detector
THAID	Theta-Aid Interaction Detector
MCA	Multiple Classification Analysis
NA	Multivariate Nominal Analysis
FACTAN	Factor analysis
MDSCAL	Non-metric multidimensional scaling
FCOMP	Factor comparison
COMPARE	Configuration comparison
CAP	Configuration analysis package
FSCORE	Factor scoring
GSCORE	Guttman scaling and scoring
CLUSTER	Cluster analysis
HICLUSTER	Hierarchical cluster analysis

Wherever appropriate, these programs handle missing data, allow for sampling weights as specified by the user, and output standard OSIRIS data files or matrices for easy use with other OSIRIS programs.

System Architecture

OSIRIS III is written predominantly in FORTRAN IV, with roughly 70,000 lines of source code across the 60 programs in the system. Most of the programs rely heavily on an extensive subroutine library for input, output, setup processing, character manipulation, conversion, and so forth. The subroutine library is about 13,000 lines of code and is a mixture of Assembler and FORTRAN IV. Each of the programs in OSIRIS operates essentially on a stand-alone basis, rather than being an overlay within a larger system. A Monitor is incorporated to aid users with JCL and job flow, but it requires minimal memory. All but one program (the larger of two regression programs, identical except for adding two-stage least squares) runs within 104 K memory under VS1 or OS/MFT on IBM 360's and 370's, and a region of about 130 K under OS/MVT or VS2. The requirements of the CDC 6X00 and Univac 11XX versions are reported to be of comparable relative magnitude.

A full system for IBM sites requires about 1200 tracks of disk, where tracks are 7294 bytes per track. Because of the stand-alone nature of each program, low-use portions of the system may be deleted from the library to substantially reduce the disk requirements when necessary. The standard job control procedures employed with the OSIRIS Monitor require another 1800 tracks of disk to provide programs with intermediate file work space, but again this can be modified to suit local requirements.

Because OSIRIS is not a single, tightly integrated system, users may add their own programs or modify their own copies of OSIRIS source code with relative ease. The subroutine library may be used to facilitate input and output of standard OSIRIS files, setup processing, and other functions, so that a user with reasonable FORTRAN experience may provide valuable augmentations to the OSIRIS capabilities.

The somewhat limited amount of information that has been generated on execution costs of OSIRIS compared with other common packages suggests that OSIRIS is typically toward the lowest-cost end of the rankings. However, it is also evident that the rankings are quite dependent on the particular billing algorithm employed by the computing installation, so that generalizations are difficult to make. Another common concern about the technical aspects of a statistical system is numerical accuracy. OSIRIS handles the known tests properly, providing either the correct values or stating that computational problems have occurred. In more general terms of freedom from all forms of "bugs" the current, release 2 OSIRIS is quite stable, being the product of evolution over many years and many hundreds of thousands of user runs.

Documentation

An extensive set of documentation is available for OSIRIS III. The basic materials are found in the six-volume OSIRIS III User's Manual (each volume available separately).

OSIRIS III, Vol 1: System and Program Description.

A thorough description of each program and overall package. The write-up for each program includes a general description, uses, functional relations to other programs, extended explanations of options and features, restrictions, input and output requirements, execution procedures and references. Not included: Error Messages (see Vol. 2) and Sample Printout (see Vol. 4). Revised edition, 1976. 846 pages.

OSIRIS III, Vol. 2: Error Messages.

A complete listing of program error comments and warnings is presented, including clarifying explanations. These messages are not found in other volumes of the OSIRIS III manual. 1973, 212 pages.

OSIRIS III, Vol. 3: Summary of Control Cards.

This volume summarizes for each program the JCL statements, the OSIRIS Monitor control cards, the program control cards and the program options. These sections are summaries from Volume 1 of the OSIRIS manual. 1976, 188 pages.

OSIRIS III, Vol. 4: Sample Jobs.

Printout from a sample run of each OSIRIS program with listings of any data that are generated are reproduced for this volume. 1973, 717 pages.

OSIRIS III, Vol. 5: Formulas and Statistical References.

This volume provides formulas, brief explanations and references for the statistics calculated by each program. 1974, 212 pages.

OSIRIS III, Vol. 6: Primer.

Covering only basic capabilities, this volume introduces the newcomer to the uses of computers and OSIRIS III.2. 1976, 126 pages.

Other, more technical information may be found in the following documents.

OSIRIS III Subroutine Manual. All of the subroutines in the OSIRIS III library are described in this manual. The functional characteristics are detailed as well as all entry points and arguments. This manual is useful to those wishing to modify existing OSIRIS III programs to add new programs.

OSIRIS III Implementation Guide. This guide contains the instructions for initializing and maintaining the OSIRIS III package. It also contains reproductions of the basic tests of OSIRIS along with the output of these tests.

Listed below are some monographs which may be of additional interest.

Searching for Structure, by John A. Sonquist, Elizabeth Lauh Baker, and James N. Morgan.

This report presents an approach to analysis of substantial bodies of microdata and documentation for a computer program. The new computer program - AID III - is a descendant of the original Automatic Interaction Detector program which started the application of search strategy. Revised edition, 1974. 236 pages.

Multiple Classification Analysis, A Report on a Computer Program for Multiple Regression Using Categorical Predictors, by Frank M. Andrews, James N. Morgan, John A. Sonquist, and Laura Klem.

Multiple Classification Analysis is a technique for examining the interrelationship between several predictor variables and a dependent variable within the context of an additive model. Revised edition, 1974. 105 pages.

Multivariate Nominal Scale Analysis, A Report on a New Analysis Technique and a Computer Program, by Frank M. Andrews and Robert C. Messenger.

This monograph describes a powerful new technique for conducting multivariate analyses of categorical dependent variables, applying an additive analytic model. It is uniquely useful in exploring the interrelationships of theoretical concepts involving categorical dependent variables and substantial numbers of independent variables at various levels of measurement. 1973. 114 pages.

THAID, A Sequential Analysis Program for the Analysis of Nominal Scale Dependent Variables, by James N. Morgan and Robert C. Messenger.

This monograph describes a recently developed technique for conducting multivariate analysis of categorical dependent variables. THAID describes a searching process which provides an efficient and effective means for sorting through a variety of analytic models to find the one most able to produce useful predictions. The program searches for subgroups that differ maximally as to their distribution; it assumes neither additivity nor linearity, so requires substantial samples of 1,000 or more cases. 1973. 98 pages.

OSIRIS: Architecture and Design, by Judith Rattenbury and Neal Van Eck.

This monograph provides technical documentation for the benefit of those involved with the writing and modification of the OSIRIS package of computer programs. 1973. 315 pages.

Data Processing in the Social Sciences and OSIRIS, by Judith Rattenbury and Paula Pelletier.

This monograph is intended to guide researchers in the field of social science (or their assistants) through all the stages necessary for processing data with a computer. It introduces the basic components of computers and the different kinds of software necessary for using a computer and then discusses types of data and some of the preliminary data collection phases prior to computer processing. Sample setups and computer output for all the major data processing steps are included in the appendices.

A Guide for Selecting Statistical Techniques for Analyzing Social Science Data, by Frank M. Andrews, Laura Klem, Terrence N. Davidson, Patrick M. O'Malley, and Willard L. Rodgers.

The Guide is intended to be useful to social scientists, data analysts, and graduate students who already have some knowledge of social science statistics. It presents a systematic but highly condensed overview of over 100 currently used statistics and statistical techniques and their uses. The core of the Guide - a "decision tree" - consists of 16 pages of sequential questions and answers which lead the user to the appropriate technique. 1974, third printing 1976. 38 pages.

Order forms and prices for the above publications may be obtained from:

ISR OSIRIS Distribution
P.O. Box 1248
Room 4250
Ann Arbor, Michigan 48106
Phone: (313) 764-6554

Availability

The distribution of OSIRIS III by ISR is intended as a self-supporting service to the user community. The Institute does not attempt to act as a vendor presenting a software product, and views OSIRIS primarily as an internal computing resource that is shared with others.

Those institutions which are members of the Inter-university Consortium for Political and Social Research (ICPSR) are offered the lowest rates for OSIRIS and some additional assistance because of the support history of this organization.

Because of current cost pressures, the prices given below are guaranteed only until January 1, 1977. After that data, increases may be necessary.

Price Schedule

I. OSIRIS III, release 2, Full System

	New Users	Previous * Users
ICPSR installations	$300	$150
Other academic & governmental installations	$750	$300
Commercial installations	$1000	$500

II. Subsets of OSIRIS III, release 2 (Costs are fixed for all classes of users.)

Package containing load modules of AID, MCA, MNA, THAID only (for OS/360 use)	$300
Source code for any individual program	$75
Source code for subroutine library	$300

* Users who purchased previous versions of OSIRIS Full system.

Further information and order forms are available from:

ISR OSIRIS Distribution
Room 4250
P.O. Box 1248
Ann Arbor, Michigan 48106
Phone: (313) 764-6554

Information on other versions of OSIRIS may be obtained from the following:

for CDC 6X00:
Universitat zu Koln
Rechenzentrum
OSIRIS Distribution
D-5000 Koln 41
Robert Koch Str. 10
West Germany

for UNIVAC 11XX:
Fred Mau
CHI Corporation
11000 Cedar Avenue
Cleveland, Ohio 44106

for DEC-10
Mr. Dick Houchard
Computing Center
Western Michigan University
Kalamazoo, Michigan 49001

Future Versions

Several changes in OSIRIS are under development at the ISR. These include a simple interactive setup processor, redesign of program internals to improve efficiency and incorporate dynamic memory allocation, improved dictionary structure and expanded data descrip-

provements. Perhaps the greatest

hierarchical file structure capability. Data collections for which a tree structure or hierarchical relationship exists between groups of variables - such as between parents, children, medical incidents, and the like - may be assigned a logical structure and stored in blocked, variable-length records in a sequential file. When such files subsequently are read by an OSIRIS program, specifications may be given by the user to create what is logically an input flat file; that is, a file without hierarchy. The user specifies which level of hierarchy forms the unit of analysis (i.e., triggers the creation of the data case passed to the calling program) and specifies how groups at other logical levels are to be placed within the record being generated. The reason for this approach is that all the existing statistical procedures expect to operate on logically flat data, and thus an ability to dynamically map from hierarchical to flat structure is required. Additional capabilities for restructuring or extending the hierarchy are also available to the user.

The timing for the release of these new capabilities to user installations outside ISR cannot be stated with certainty; however, it will undoubtedly be at least a year before distribution begins. In any case, full compatibility with existing standard OSIRIS data files will be maintained.

Some individual programs which augment the capabilities of OSIRIS, but which are not integrated into the current package, may be obtained from the Institute. One such collection of programs, known as SRCLIB, includes routines for computing sampling errors, a maximum likelihood regression, the MINISSA multidimensional scaling program, a univariate and bivariate frequency program emphasizing low computational costs, and a number of others. For further information contact:

SRCLIB Distribution
Room 106
Institute for Social Research
P.O. Box 1248
Ann Arbor, Michigan 48106

Other References

Several articles exist which provide further information and points of view about particular aspects of available packages, including OSIRIS.

Francis, I., Sherman, S.P., and Heiberger, R.M., "Languages and Programs for Tabulating Data from Surveys," Proceedings, Computer Science and Statistics: Ninth Annual Symposium on the Interface, Cambridge, 1976.

Robinovitz, S.H., "The OSIRIS Data Management and Analysis System," Proceedings of SHARE XLVI, San Francisco, February, 1976.

Roistacher, R.I., "A General Consistency Check Procedure for Machine-Readable Data," Sociological Methods and Research, Vol. 4, No. 3, February, 1976, pp. 301-320.

Slysz, W.D., "An Evaluation of Statistical Software in the Social Sciences," Communications of the ACM, Vol. 17, No. 6, June, 1974, pp. 326-332.

Slysz, W.D., "Performance Differences in Social Science Statistical Software," in Facts and Futures, Proceedings of the EDUCOM Fall 1973 Conferences, Princeton, EDUCOM, 1974, pp. 235-243.