APPENDIX 3A

IRIS Data Management Center

Program Plan

Table of Contents

APPENDIX 3A:
IRIS DATA MANAGEMENT CENTER PROGRAM PLAN

1. INTRODUCTION AND RATIONALE

The function of the Data Management Center (DMC) is to assure that the high-quality data sets generated by the PASSCAL and GSN data collection systems together with appropriate data sub-sets obtained from other U.S. organizations and other countries are made available to the seismological user community in a timely manner and in a tractable form. The goal is to make it possible for a researcher to concentrate primarily on the analysis and interpretation of data rather than the assembly of usable data sets. Thus, effective dissemination of data to research seismologists is the single priority that drives the overall Data Management Center requirements.

The guidelines used in establishing the functional requirements of the proposed DMC included in this proposal come from the discussions of the scientific data needs that appear in the accompanying GSN and PASSCAL plans and from the recommendations in the 1983 National Academy of Sciences' report entitled "Effective Use of Earthquake Data" (Appendix 3B). This proposal also embraces the principles of successful scientific data management and the philosophy of operating Scientific Data Management Units as set forth by the Space Science Board Committee on Data Management and Computation in its 1982 National Academy of Sciences report entitled "Data Management and Computation, Volume 1: Issues and Recommendations." In the context of that report, the IRIS Data Management Center would be a Disciplinary Data Management Unit rather than a Principal Investigator or Project Data Management Unit. Thus, the proposed IRIS Data Management Center will serve the seismological user community through the dissemination of both IRIS generated dat, and certain selected seismological data collected by other U.S. organizations and other countries. It will serve the role of a national data center for digital seismological studies in the sense recommended in the N.A.S. report (Appendix 3B), but it is not intended to provide a comprehensive archival source of all available digitial seismic data that are being collected nationally or internationally.

The Data Management Center will be required to handle and distribute digital data that spans the spectrum from fixed network recordings (GSN and sub-sets form other U.S. and foreign networks) to controlled-source portable array recordings generated by PASSCAL experiments and selected data from other U.S. or foreign experiments. In some respects, the mixed mode of fixed and portable array recording of earthquakes (or explosions) over a protracted time interval will present the most challenging data management problems. Close communication and co-ordination with the GSN and PASSCAL Standing Committees will be essential from the beginning of the IRIS program to assure that the scientific needs that prompted the GSN and PASSCAL initiatives are met. Feedback from users will be strongly encouraged to help assure effective dissemination of the data.

The extraordinary advances in all aspects of digital computing technology over the past decade have provided the technical resources to handle the data volumes anticipated under the IRIS programs. As in most large data gathering efforts, the real challenge is effective data management. The rationale for structuring the proposed Data Management Center is that effective distribution of data to users in desired forms for analysis and interpretation must drive the system. Provisions for incorporating further advances in computer hardware and software capabilities, as needed, are also important in structuring the Data Management Center.

The approach proposed is to develop in the first 18 months the detailed design requirements for the Data Management Center that will be needed to accommodate the schedule of data collection anticipated under the GSN and PASSCAL science plans. Allowing time for procurement and installation, the resulting facility would become operational during the third year.

In addition, early data distribution and data management experiments are planned for the first two years using existing facilities developed for other purposes. These experiments will serve to provide some useful prototype IRIS data to the seismological community immediately and to get important feedback on which modes of distribution are likely to be most effective.

The sections that follow treat the functional requirements of the Data Management Center, the proposed organizational structure, the implementation plan, and the schedule and budget for a 10-year interval.

2. FUNCTIONAL REQUIREMENTS

2.1. Parameter Data Base

The bulk of the data to be held by the DMC consists of digitized waveforms, although a smaller dataset must exist to make the waveform data useable. This "Parameter Data Base" must maintain a comprehensive data directory that provides an index to data holdings at the center. The system must provide rapid and easy access to bulletin hypocenters and associated phase data, and to other derived data such as focal mechanisms and moment tensor solutions.

The management system should allow both casual and detailed inquiries into the location and nature of PASSCAL experiments. The parameters must include: geographic location, source/receiver heights, source types (including earthquake, explosive, vibroseis, and airgun), receiver types and sensors used. Relations for passive earthquake experiments should be capable of cross-referencing global catalogs. Finally, decimated, processed, or stacked data should be available in a reasonable period of time for preview.

2.2. Waveform Data Base

The digitized waveform data form a large and rapidly growing dataset which must be properly treated in order to remain accessible and usable. The DMC must provide large on-line and off-line storage capability (mass storage if available) for digital waveforms data from GSN and PASSCAL stations and arrays worldwide. The data base will eventually include digital data from earthquake strong motion systems and from special event datasets. The expected GSN data volume wil be about 1.0 Terabyte/yr, and for PASSCAL, about 0.4 Terabyte/yr.

2.3. Data Base Management System (DBMS)

The DMC will develop and maintain an effective relational data base management system to retrieve, on request and in a timely manner, parameter and waveform data in an integrated form, independent of its storage location. The system must provide the versatility to meet a variety of types and combinations of features specified by data users in the form of "seismological queries." In addition to this larger service, the DBMS should allow the general user to construct subsidiary data bases for specific research projects. The DMC must provide for rapid archival of data with a retrieval architecture structured to accomodate specific user needs and frequency of use.

2.4. PASSCAL Data Processing

The demands placed on the data management facility by PASSCAL will be significant. Although the data will generally be provided in a standard, e.g. SEG Y, format, the source association will generally be incomplete and must be completed at the Center under the supervision of the Principal Investigators. Earthquake data present the greatest burden in that the field computers will not be capable of performing even rudimentary event associations except for very limited deployments. These datasets will also be quite large (200 Gbtes over a 300 day period) and will significantly tax the storage and bus speeds available in the data management facility. The planning model currently will provide on the order of 410 Gigabytes each year when the PASSCAL data system is fully operational. Further details concerning the nature and volume of this data can be found in the PASSCAL Program Plan.

2.5. GSN Data Processing

The requirements of the data processing system will be to:

1) provide a capability to perform automatic or interactive event detection, phase association, event location and preprocessing of waveform data.
2) provide means to perform real-time data manipulation.
3) implement and maintain selected analysis techniques such as those needed to provide a near-real-time capability to determine source characteristics of large earthquakes for purposes such as warning (e.g., tsunami), damage assessment, and -deployment of special instrumentation in the epicentral area.

2.6. Communications

The objective here is to maintain interfaces in order to provide the Center access to remote data bases and to permit remote, user-friendly, access to the Center's data bases. A remote user should have access to the DBMS and should be able to preview data on a limited basis. The user should, in addition, be able to dispose arbitrary parameter and waveform data to various hard copy devices. It will also be necessary to develop and maintain a satellite communications interface to handle the data stream broadcast from GSN stations.

2.7. Distribution

There will be distribution of data to users by a variety of standard techniques and formats to include: remote data base access by a "seismic work station", network-day, network event and special event types in standard format, common source or common mid-point gathers, and analog products such as paper recordings, fiche, and videodisk.

2.8. User Services

The system will provide the capability to quicklook and browse for waveform data. A Principal Investigator should have high priority access to the database computer for processing data resulting from recent field experiments. There will be a provision for preprocessing of data (filtering, spectra, record sections, etc.) and for a variety of waveform graphical display choices. It will be ncessary to permit limited access to computers for intensive data manipulations and preliminary analyses of data. A user's guide, user software, and experimental or tutorial data sets will be developed and maintained. Periodic training sessions for new users and a visiting scientist program will be offered. Other tasks will include the dissemination of information about the Center's capabilities and available data bases as well as distribution of software. Periodic review sessions or special symposia will be organized to discuss results and capabilities and to identity additional user needs.

2.9. International Cooperation

The high quality digital data sets generated by PASSCAL and GSN data collection systems will constitute a revolutionary development in seismic instrumentation. At the same time it is envisaged that initiatives in other countries will also contribute significantly to the global data set, and that, with suitable agreements, such data will become available to U.S. seismologists. Even in the U.S., it is unlikely that IRIS will come to monopolize the collection of seismic data. The DMC will take the initiative in acquiring data from both foreign and U.S. networks for its collection. An important aspect of its operation will be the maintenance of a complete archive and a complete catalog of its holdings. Under suitable agreements, its archive could be made available to Data Centers in other countries, though its primary responsibility would be to satisfy requests for data from U.S. researchers. Its data catalog would also be circulated to assist in formulating data requests.

2.10. Development

We anticipate that many facilities for serving scientists' needs will perform in a rather rudimentary fashion during the early part of the program. It is essential that ongoing development of both hardware and software be pursued aggressively in order that the system respond to the increasing expectations of the community. New technology and software must be implemented as they become available to improve the Center's capabilities to provide data services and products.

3. ORGANIZATIONAL STRUCTURE FOR THE DATA CENTER

The plan for the organizational structure of the data center follows these basic guidelines:

1) The number of permanent IRIS employees shall be minimized, and these employees will be largely involved in managerial activities and coordination with IRIS standing committees and with the user community.

2) As a consequence, most operations of the data center will be performed by contractors on a competitive basis.

3) The structure should be simple and flexible, and, in particular, allow for a smooth transition through the various developmental stages of the data center.

Our present view of the organizational structure is best conveyed with a simple diagram.

The program manager will be responsible for the proper performance of all the data center activities, as defined by the Standing Committee. Clearly, this is a full-time task which also requires a modest support staff (for clerical work, monitoring of contracts, etc.). The program manager would frequently participate in the Data Management Committee meetings, meet with the IRIS President and other IRIS standing committees as appropriate, and be a member of sub-committees of the Data Management Standing Committee. The recommendations of the detailed data center design studies that address the organization structure will be used to formulate the specific responsibilities of the program manager and his staff versus those of individual contractors.

A program manager must be hired as soon as possible. We need a full-time person to be responsible for the initial data center activities. In particular, the program manager will be responsible for monitoring the design requirements contract(s) and for coordinating the data distribution experiments that are planned for the first two years.

4. DESIGN AND IMPLEMENTATION

We believe that the size and diverisity of the data sets that will be generated by GSN and PASSCAL, as well as the requirement for wide distribution of this data, present problems of a scale beyond any current models of data handling within the seismological community. Development of this system must be undertaken with a full appreciation of the enormous quantities of data to be accessed at any time and of the high bus speeds required for this task. The planned and predictable progress in computer and storage technology will obviously result in a substantially enhanced capacity and reduced costs over the several years that will elapse between the planning and implementation phases of this project. Furthermore, the growing trend toward multiple processors in both super-mini and mainframe computers may allow the data management system to be more simply realized than current architectures will permit. Even more radically new and unproven systems using a parallel processor archictecture might provide yet other alternatives in the future.

As a result of these considerations, our strategy involves a two-stage design process, in which we first will determine the functional requirements of an effective data management facility before undertaking a more detailed design and system specification that will enable these objectives to be met. With the help of experts from both systems design and computer science interacting with seismologists and managers of existing seismological data services, we hope to enter the design specification phase with a set of definitions and functional specifications of sufficient clarity and detail to insure a successful final product. We expect to accomplish both stages of the design process by contracting with commericial entities.

4.1. Schedule

Data flow is expected in FY85 and 86 with interim experiments of both PASSCAL and INS. It will be handled primarily by the cooperating institutions that generate the data, but will also flow through the prototype data facility insofar as that is possible. Beginning in FY 87 there will be data from a few of the new GSN and PASSCAL stations, and the data center will begin to assume responsibility for archiving and dissemination at that time. The initial volume of data will be small and will gradually increase to nearly its full design value by FY90. The schedule of design and implementation presented here reflects the projections of data flow described in the GSN and PASSCAL program plans.

The development of hardware and software systems required to bring the Data Management Center up to full operational capability will be acquired in a series of three-staged procurements beginning in 1985 and continuing through 1988, with completion anticipated in early 1989. In the first stage, the preliminary functional system specifications will be developed and an initial contractor selected by a DMC technical committee charged with this responsibility. This stage, during the first half of 1985, will draw heavily on the prototype system developed in parallel by the PASSCAL Data Management Project discussed above. This initial contract is expected to provide a detailed definition of the DMC requirements by the end of 1985. These requirements will form the basis for a solicitation for a second contract, which will in turn provide a detailed specification document by the end of 1986 for a fully operational hardware and software system. A contract for final system design and implementation is expected to be in place by early 1987. Implementation and installation is carried out in the third stage, beginning in mid-1987, with full hardware and software capability expected by the end of 1988. During subsequent years, the development budget primarily reflects software maintenance expenses as well as those required to retain 'state of the art' hardware capability.

4.2. Budget Considerations

While reliable figures for implementation costs will not be available until sometime in 1987 when the final design document is nearing completion, reasonable budget estimates can be obtained from a consideration of the hardware and software costs of existing systems that implement portions of the capabilities required for IRIS data management. Such comparisons support the conclusion that our needs will probably be adequately served by a cluster of high capability minicomputers, or perhaps several small main-frames with between 10 and 50 gigabytes of direct-access storage. With existing technology, from 10 to 50 1-Gbyte Winchester disk drives would be satisfactory. While this configuration can be used to project reasonable budget estimates, substantial improvements in both price and performance will undoubtedly occur before the design document is finalized.

PASSCAL plans to produce some very large data sets, perhaps a fraction of a terabyte, for-three dimensional imaging problems. Such data sets will require sort and gather processing that will be beyond the capability of the PASSCAL field computers, and thus will either have to be handled by the IRIS Data Management Center or contracted out to commercial services. A final decision cannot be made until system specifications are available, so that a comparison of costs can be made between enhancing local computational resources and performing front-end data reduction off-site on a contract basis.

5. START-UP DATA MANAGEMENT AND DISTRIBUTION EXPERIMENTS

By taking advantage of existing data center facilities developed by other organizations for other purposes, present and anticipated global network data set, and data sets that will be collected in PASSCAL-related field deployment s during FY 85-86, important prototype IRIS data management and data distribution experiments can be carried out. These experiments are intended to identify problem areas to get user input to the Data Management Center design via feedback from using the prototype digital data sets and to provide a limited amount of digital data needed for ongoing seismological research efforts at IRIS member institutions. While distribution of these prototype data sets would not be limited to IRIS member institutions, their needs would have high priority.

In the following sections are described start-up activities using existing facilities of other organizations, as indicated in the discussion of each activity.

5.1. GSN-Data Experiments

5.1.1. IASPEI International Data Set

The IASPEI Commission on Practice is organizing a special session on the Analysis of Selected Earthquakes at the 1985 General Assembly in Tokyo. The goal of the session is to provide a focus on modern scientific practice in the areas of earthquake quantification, data exchange, digital seismology and algorithms. Participants have agreed to analyze a selection of fifty-one recent well-recorded earthquakes worldwide, representing a range of sizes, locations and focal depths, and source characteristics. To facilitate these analyses, participants will be provided data in digital format, software to read the data, and other related materials from a variety of sources.

We recognize that this special IASPEI data set presents a unique opportunity to test a variety of procedures and formats for data dissemination of the type that might see future implementation at the Center. The USGS plans to test and expand its digital data distribution services for this dataset by means at its disposal. These include distribution of network-event tapes, hard copy waveform catalogs, and event parameters and associated phase data on magnetic tape.

More advanced distribution methods will be implemented by IRIS staff at the DARPA Center for Seismic Studies and elsewhere to permit remote access and manipulation of a special purpose data base constructed from the IASPEI data set. Construction of this special data base will provide an opportunity to integrate international event waveform recordings input in different formats and to form experimental displays (e.g. record sections; 3-component azimuthal sections; Z, R, and T versus distance and azimuth; pseudo-3 dimensional plots, and other innovative means of display). Digital sub-sets of this date base can be provided to users in a standard format, but probably will be limited to user-specified time or velocity windows.

5.1.2. Other Selected Event Data

From the currently available GDSN network day tapes on the GDSN network and event tapes for m_b 5.5 events, a limited amount of digital waveform distribution can be accommodated based on user requests. These requests would be constrained to user-specified time or velocity windows for a specified set of stations on the entire network. A choice of modes of distribution would be provided (e.g. written request or remote terminal request for data to be provided on magnetic tape, or transmission of waveform data via high baud rate line to a user work station).

This activity must be limited in scope because it depends on shared use of facilities of other organizations, such as the DARPA Center for Seismic Studies (CSS). Therefore, at most, one modest data request from each IRIS member institution can be considered.

5.2. PASSCAL Data Experiments

The PASSCAL instrumentation has been specified to be applicable to a wide range of experiments. One class of experiments, those using artificial sources such as explosions or vibrators, is analogous to current oil industry seismic exploration techniques. While data volumes are very large, the source times and locations are known; therefore, the data collection and preparation stages, though I/O intensive, can be modelled after those methods developed by the industry over the last few decades. The second class of experiments, involving natural sources, poses further processing difficulties. For these experiments, the formidable front-end task of seismic event detection, association, and windowing is added prior to the creation.of some event oriented data structure. There already exists a great deal of experience and capability in this area. The problem is identical to that already addressed by various local- and regional-scale digital seismic networks. Solutions of various stages of completeness have been produced by the USGS, DARPA, DOE national laboratories, and various universities. The further stage of processing proposed for the data center, which involves the manipulation and analysis of the waveform data, has also been addressed by these same institutions.

The existing data management working group of PASSCAL is currently in the process of evaluating and ,where appropriate, integrating these capabilities. From this effort will emerge a prototype data handling system which will be capable of processing data from some hundreds of stations. Clearly, any further specification of the software needed for the datacenter will be made much more valid if it builds upon a careful evaluation of the existing expertise and capabilities.

5.3. Evaluation of Prototype Datasets

In addition to assembling the prototype processing software, a series of data set assembly experiments are proposed. The more immediate of these involves data sets using natural sources and making use of a collection of a few hundred seismic recorders provided by various institutions; eg. USGS, LLNL, LBL, University of Wisconsin, and others. The target area for this experiment is the Long Valley, California, region. The data set will provide some additional difficulties which will not be present in later PASSCAL data in that each institution has a different format for data recording. Nonetheless, it is proposed to collect some hundreds of channels of data for a number of seismic events in this area, and to undertake the assembly of a prototype data set which can be distributed to the IRIS membership for evaluation. The data collection efforts are expected to be funded from ongoing programmatic projects at each institution, and the assembly of the dataset will be done using the prototype software being integrated by the PASSCAL data management working group. The facilities for the work of assembling the data set are available at LLNL, LBL, USGS Menlo Park, and USGS Pasadena. The PASSCAL data management working group will be responsible for the assembly of the prototype dataset.

A second data management experiment (DSS) is also proposed which will make use of a large collection of controlled source digital seismic data. Desireable datasets include that produced by the non-spatially aliased DSS line of about 2000 km, using NTS shots as sources and leased commercial oil exploration equipment for recording. This experiment, currently under consideration for a possible FY 1986 deployment, would produce data from 1000 to several thousand source/receiver pairs. The data collection and initial assembly would be contracted to a commerical exploration firm. This would allow the evaluation of such arrangements for future PASSCAL experiments. The assembly into an IRIS data set and dissemination to the membership for evaluation would, again, be the responsiblity of the PASSCAL data managment working group. Another desirable dataset for such evaluation is that proposed for collection in the vicinity of the proposed Atlantic Coast deep drilling site. In either case, the data collection task is expected to be funded by the Prinicpal Investigators involved, but the formatting into an IRIS data set and interaction with IRIS members about the suitability of the datasets will be conducted by PASSCAL.

The creation of a prototype data management system and prototype datasets is intended to make the fullest use of existing capabilities in the definition of new full-scale data center, to provide the IRIS membership with tangible products indicative of those which will be produced routinely by the PASSCAL system, and to allow those members to provide evaluation and feedback on data set organization, quality, etc., before the implementation of the full-scale data center.

5.4. Hard-Copy Products Evaluation

In order to acquaint the user community with the types of digital data that are available and those that will be forthcoming under the IRIS programs, several types of hard-copy media and several types of display will be evaluated. The principal objective is to determine the usefulness of hard-copy displays as a means for users to browse for useful data sets for their research and to familiarize themselves with the data sets that will be available in the IRIS Data Center.

Both PASSCAL and GSN data sets will be used to generate hard-copy displays on at least three different media: paper copies, micro-fiche, and videodisk. The display formats to be evaluated will include event record-sections, event azimuth sections, and perspective plots.

IRIS DATA MANAGEMENT CENTER 10 YEAR BUDGET PLAN

	FY85	FY86	FY87	FY88	FY89	FY90-94
STAFF	280	390	500	500	500	500
START-UP EXPERIMENTS	510	410	100			
SYSTEM DESIGN	350	500	100			
IMPLEMENTATION						
Hardware procurement			3000	2000	1500	
Hardware upgrade/replacement						1500
Initial software development			3000	2000	1000	
Ongoing software R & D and upgrade		500	1500			
Operations and maintenance			1000	2000	2000	2000
USER SUPPORT SERVICES			400	1000	1000	1000
COMMUNICATIONS	150	150	300	700	1000	1000
STANDING COMMITTEE SUPPORT (Including Technical Committees and Workshops)	150	150	150	150	150	150
TOTALS	1240	1800	8550	8350	7650	7650

DMC INITIAL 2-YEAR BUDGET PLAN

	FY85	FY86
STAFF		
Program Scientist/Manager	150	150
Seismologist/Engineer	100	100
Systems engineer		100
Office/clerical support	30	40
SYSTEM DESIGN		350
START-UP EXPERIMENTS		
PASSCAL Prototype		
Long Valley & D.S.S. Software integration and interchange	*	*
Preparation of data for distribution to users	50	100
GSN Prototype IASPEI and GSE international data exchange experiments	10	60
GSN skeletal data distribution system		100
Hard-copy distribution testing (fiche, video, paper)	50	50
Data Center Prototype Interim facility/Headquarters	50	75
Computer use and equipment rental	150	175
COMMUNICATIONS		
Leased lines and communications node rental	30	100
Terminals and workstations	70	50
TOTAL	1240	1800