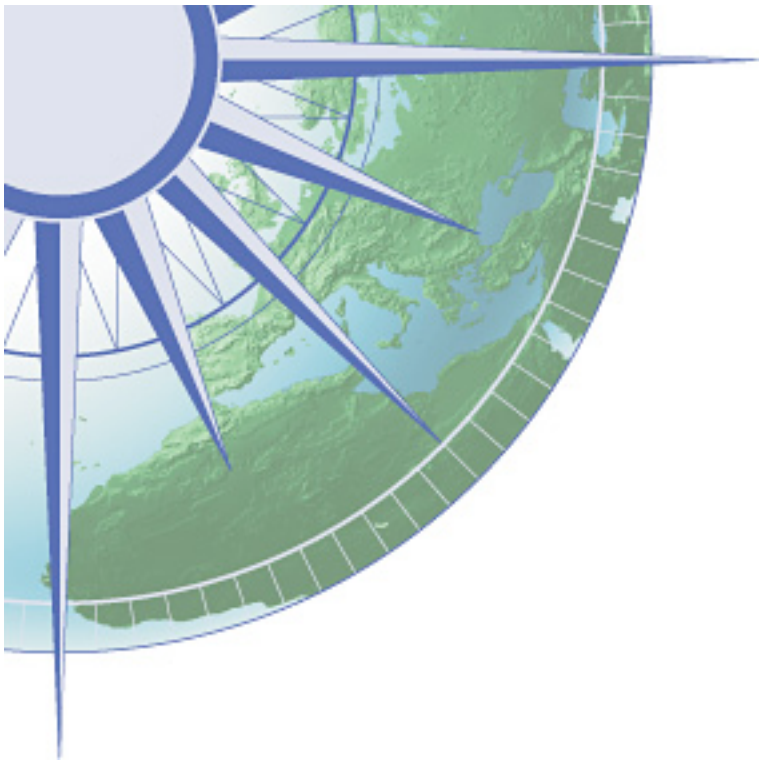

Maintaining and Improving Access to Earth Science Data

A Strategic Plan

for the

IRIS Data Management System

of the

Incorporated Research Institutions for Seismology

November 2003 Version 1.0

I. Introduction

To maintain vitality and relevance, the various components of an organization need to be periodically reviewed in one of a variety of ways. The Data Management System (DMS) of the Incorporated Research Institutions for Seismology (IRIS) began this process with an internal review concluded in 2001 by then Chair of the DMS Standing Committee (DMSSC), Alan Levander. This resulted in the DMS Internal Review document (Levander, 2001). This review recommended that instead of conducting an External Review of the DMS at that time, it would be more beneficial to develop a Strategic Plan for the IRIS DMS that could guide the future course of this key component of IRIS.

This DMS Strategic Plan should serve as the basis for the periodic review of the DMS component of IRIS. A reasonable timeframe for the review would be roughly every three years, coinciding with the rotation of the chair of the DMS Standing Committee.

This plan was developed using a streamlined process. Input was solicited from the oversight committee structure of the IRIS Consortium including 1) the IRIS Data Management System Standing Committee (DMSSC), 2) IRIS Staff and 3) the Executive Committee of the IRIS Board of Directors (EXCOM). The DMSSC and EXCOM reviewed and approved this strategic plan. In addition to direct community input through the various committees, staff of the IRIS Data Management Center (DMC) in Seattle, the USGS Albuquerque Seismological Laboratory (ASL) and the University of California San Diego IDA Data Collection Center (DCC) contributed to this plan.

We have evaluated the strengths of our system and learned during the process of developing this plan. Our greatest strength is in our ability to directly meet the needs of the national and international seismological communities. We reiterate this as a fundamental tenet of the DMS as we move forward. Significantly, the planning process also identified the need to move out of the confines of seismology only and to extend some of the techniques and data management principles we have mastered to other closely related geophysical disciplines.

We intend to move forward into these new disciplines through close and careful dialogue with leaders in these fields and to develop data management techniques that encourage the grass roots input to define the characteristics needed while still retaining the overall structure and standardization that have

allowed the IRIS DMS to be viewed as a leader in data management.

Tim Ahern,
Program Manager,
IRIS Data Management System

II. Mission and Vision

The Mission of the IRIS Data Management System

To provide reliable and efficient access to high quality seismological and related geophysical data, generated by IRIS and its domestic and international partners, and to enable all parties interested in using these data to do so in a straightforward and efficient manner.

This mission statement builds upon the strengths that have been developed over more than a decade and identifies new, shorter-term directions that can serve to guide the DMS.

In order to fulfill this mission, the IRIS DMSSC, who played an important role in its formulation, identified some specific guidelines that will provide direction as to how we can succeed with our mission. As we anticipate available funds will not grow significantly, these guiding principles stress participation of the broader community, forging alliances with other organizations and always finding ways to improve efficiencies within the DMS.

To fulfill its mission, the IRIS DMS must:

- Strive to be highly automated
- Develop systems to insure high quality and complete datasets.
- Insure quality control is done in a cooperative manner by encouraging data users to report quality concerns to network operators.
- Develop easy-to-use data quality reporting systems.
- Collect, manage and distribute seismic data taking advantage of proven information technology in a proactive manner.
- Facilitate the efficient use of data within the DMS and by end-users of the data distributed by the DMS.

The IRIS DMS Vision

The IRIS DMS will be a leader in the archive and dissemination of seismic and related data to and from anywhere in the world. Furthermore we will be a resource for technology supporting scientific analysis of these data.

This vision serves to provide the longer-term goals for the DMS so as not to become narrowly focused upon our shorter term successes and problems but also

continually looking where we need to be several years in the future. It provides the breadth the DMS needs to continue to improve services and data availability so as not to become complacent.

III. Critical Issues and Core Strategies

The expectations of the seismological research community are always increasing and yet the core funding of IRIS and therefore the IRIS DMS remains static. Strategies that the DMS should employ to deal with this static funding include:

- Sustain core funding to maintain existing capabilities.
- In general, do not adopt new software and hardware technologies until they are proven to be viable by other commercial or research organizations. However, the DMS should always be aware of developing directions in software, hardware, and IT in order to adopt those which will serve the long term DMS needs.
- Develop proposals to secure funding from other sources to develop and utilize new techniques.
    - The NSF Information Technology Research and Cyberinfrastructure programs are examples of programs that might be used to augment the available funds for DMS development.

Research needs increasingly call for integration of data of different types whereas the IRIS DMS' expertise is highly focused on the management of seismological data.

Consistent with the above-stated DMS Mission, the DMS should investigate expanding its data management expertise into additional geophysical data types that could be linked with seismological data and do not presently have comprehensive and professionally managed collections.

The DMS should consider hosting some data types such as

- magnetotelluric data
- strain data
- Earthscope complementary datasets.

The DMS should establish tight links to other comprehensive data centers that possess long-term viability such as those at

- USGS and
- UNAVCO.

The DMS should participate in IT data integration projects such as:

- GEON and GEOINFORMATICS
- CYBER Infrastructure
- SCEC-ITR

In addition to the knowledgeable staff and software infrastructure in place at the DCCs and the

DMC, a very important asset of the IRIS DMS is the archive of well-managed seismological data and the related metadata that describes the data. The DMS should:

- Develop a formal data transcription policy that requires transcription of all observational data and its associated metadata to new media with a period of not greater than five years.
- Develop a formal disaster recovery plan that insures the protection of all data, metadata and software applications in the event of the destruction of the IRIS DMC.
- Develop high availability infrastructure at the IRIS DMC to minimize downtime.
- Protect the IRIS DMC from extended disruption in the event of major natural catastrophes.

To increase the data that are available to the seismological community the DMS should:

- Continue working with the FDSN to insure continued access to subsets of data of the FDSN membership.
- Continue to work to expand the membership of the FDSN.
- Work to make data from non-IRIS portable experiments available.
- Work with the USGS to insure that data flow from the US Regional Networks and the Advanced National Seismic System of the USGS to the IRIS DMC.
- Support short-term (<5 years) projects in select international locations to make data available from other networks through the DMC.
- Employ new but proven technologies to increase data access and ease of use of the IRIS DMC.
- Develop distributed data center connectivity tools.
- Work with the exploration industry to bring key datasets into the IRIS DMC.

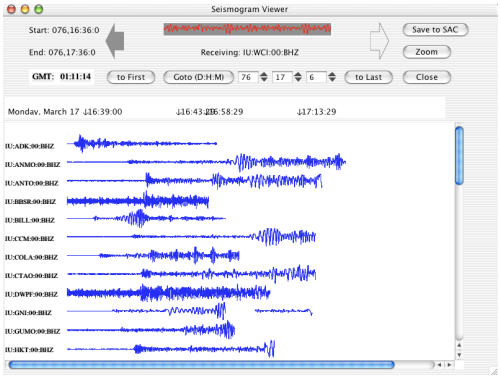

Figure 1. The Data Handling Interface (DHI), with a suite of DHI-enabled clients such as VASE that is pictured here, allow client-side applications to directly interface with metadata and waveforms held at the IRIS DMC.

Strategy for Accepting New Data Sets at the IRIS DMC

The IRIS DMC is viewed by a large segment of the Earth science community as being a desirable data center for the long term archiving and distribution of Earth science datasets. As such it is crucial for the IRIS DMS to have a well-established policy to

guide the acceptance of new data sets.

For seismological waveform data from NSF-EAR funded projects, the requirements are:

- The data must be fully documented with metadata
- The data are in one of the two formats accepted by the IRIS DMS
    - SEG-Y (Active Source)
    - SEED(Passive Source)
- The DMS Standing Committee has reviewed and approved the request to archive the new data set.

For waveform data from NSF projects outside of EAR, the requirements are:

- The data must meet all the requirements for NSF-EAR projects.
- Either the level of effort is not significant for the DMC to manage these data or the sponsoring agency will provide funds to cover the incremental costs for data management.
- A system to insure data quality is in place and reasonable. This implies data control nodes similar to the Data Collection Centers of the IRIS GSN are in place.

For waveform data from non-NSF sources such as industry data sets:

- The data are in an "acceptable format" which will be determined by the IRIS DMS on a case-by-case basis.
- The DMSSC concurs that this is an appropriate data set for the IRIS DMS to manage.
- Financial resources are available to meet the increased demands.

For other types of data:

- The DMSSC concurs that this is an appropriate data set for the IRIS DMS to manage.
- The DMC can manage the data type within identified resources (old & new) including development costs and storage demands.

Education and Outreach

The DMC supports IRIS E&O efforts in the establishment and development of a National Educational Seismic Network for K-16 science education as well as IRIS E&O Earthscope outreach efforts.

- The DMC will maintain an archive of data from educational seismograph stations that are of high enough quality for educational use, including student research projects, and provide data to national users of the Educational Seismic Network.
- The DMC will cooperate with the IRIS Education and Outreach Program in the development of web-based systems for the E&O communities.

Efficiencies

To improve efficiencies the IRIS DMS should:

- Strive to acquire as much seismological data in real time and through electronic methods as possible.
- Develop automated methods to estimate data quality and other parameters from observational data in the DMC.
- Develop a strategy to reduce the amount of data that is resent to the DMC from the GSN DCCs and the PASSCAL program.

To evaluate all DMS nodes and their importance annually, the DMS should:

- Continue the DMSSC critical review of all DMS nodes.
- Limit support of non-core nodes to not more than 5 years without formal reevaluation of the usefulness of the nodes.
- Phase out funding of non-performing or non-essential nodes.

Goals and Objectives

The IRIS DMS needs to continue to make data management of data from the IRIS GSN and the IRIS PASSCAL programs the highest priority data sets we manage.

Recognizing that IRIS alone can not install and operate all the necessary stations on a global basis, the DMS must work closely with the international community through the FDSN, and regional networks world-wide to augment its holdings in a strategic manner. These needs result in the following objectives for the primary components of the DMS.

ASL and IDA DCCs

Since the number of GSN stations is not increasing, the two DCCs should continue to perform their primary functions without major expansion.

The DCCs should continue to improve efficiency and, where possible, take advantage of developments taking place elsewhere. Some specific activities include:

- The DCCs and the DMC should participate fully in the development of the DMS Automated Quality Assurance effort of the IRIS DMS.
- The DCCs should adopt an XML-based station history system and work with the network operators to encourage adoption for field use as well.
- Station documentation that can be of use to end-users of the data should be made available through the XML station book method.
- The DCCs should continue to improve quality control procedures and fold them into the Quality Assurance System being developed by the IRIS DMS.

- The DCCs should continue moving toward TCP/IP connectivity to all stations to improve data availability and reduce latency
- The DCCs should attempt to have quality-controlled data at the DMC within 3 days of their recording in the case of telelmetered stations.

IRIS DMC

The DMC should strive to be the place where seismologists and other Earth scientists look first for their data needs.

- Attempt to service most requests for data in no more than 3 hours.
- Continue developing methods to receive data electronically
  - By real time methods when possible
  - By standard protocols such as ftp where appropriate
- Insure that data can be received, processed, archived, and distributed to end users, in an entirely automated fashion.
- Develop methods to automatically update metadata and synchronize data holdings with other centers.
- Develop liaisons with international groups to make data from portable instrumentation programs in other countries available to the research community.
- Continue the development of distributed data center concepts such as
  - NetDC, networked data centers
  - Data Handling Interface (DHI) technologies which inherently support distributed data centers.

IV. Management and Operation: Goals and Objectives

The DMS must retain its close links to the research community. The system of governance now in place is highly effective and we should not envision changes to this system. The DMSSC, COCOM and EXCOM insure that decisions that affect the DMS are well reviewed and done as part of a very open process, with policy decisions and priorities set by the oversight committees.

The largest costs in the DMS are those for staff. For this reason our goal is to maintain as small a staff as possible, while still meeting the community's expectations in terms of data availability and response time. We should do this by insuring that DMS information technology is highly efficient and scalable. It is our belief that software and computing technologies used by the DMC and DCCs should always be highly efficient rather than to add staff to resolve failures in the information technology being used.

The key to a successful DMS lies with the people that are involved in its operation. The staff at the DCCs is under the control of the USGS and UCSD. These groups remain responsive to the IRIS committee structure.

Staff at the IRIS DMC is IRIS employees. Close linkages are maintained with research faculty at the University of Washington. The IRIS DMS Program Manager does direct management of the IRIS DMC staff.

This system has proven to be effective and responsive to the desires of the IRIS advisory committees. We do not see a need for major changes in the structure at the IRIS DMC and the two DCCs. At all times, responsiveness to the recommendations made by the IRIS committee structure should be maintained.

Overall DMS program management is the responsibility of the IRIS DMS Program Manager. The Program Manager interacts on an as-needed basis with the two DCCs and many informal communication lines exist between DCC and DMC staff as needed.

One element of the effective management of the DMS program lies in the annual review process of all nodes of the DMS. Each year, all nodes must provide a response to a formal Request for Information that is then reviewed by the IRIS DMS Program Manager and the DMS Standing Committee. In the past this process has resulted in reduction of support for some nodes, a change in direction of other nodes and increased

activities at other locations. The annual procedure of DMS Node review is important and should continue.

Another important forum for coordinating the activities within the IRIS community is a meeting of the people involved in the day-to-day management of seismological data. The IRIS DMS sponsors a DMC-DCC meeting where key individuals from the IRIS DMC, ASL DCC, IDA DCC, PASSCAL PIC, and some US Regional networks meet to discuss issues of significance. This is an effective forum and should continue.

The DMS staff is encouraged to meet IRIS community members at opportunities such as the IRIS annual or AGU meetings.

V. References

Levander, Alan. 2001. "Data Management System, Self-Study Report 2000 – 2001". http://www.iris.washington.edu/about/DMC/DMSSC/DMS_SelfStudy.pdf

Minster, J.B. and Goff, R.C. 1986. "Strategies for the Design of a Data Management System". Science Horizons, Inc.

TASC Inc. "Design Study for the IRIS Data Management Center", 1987. The Analytical Sciences Corporation.

Simpson, D., Prescott, W. and Zoback, M. "Earthscope: Acquisition, Construction, Integration and Facility Management". 2003.

Appendix I.
Organization Profile and History (September, 2003)

IRIS was formed in 1984 by key institutions in the United States seismological community. While improving the community's ability to field and operate permanent and temporary seismic recording systems motivated the creators of IRIS, they identified the need to have a significant data management component within the structure of IRIS.

Two initial studies guided the development of the IRIS DMC. "Strategies for the Design of the IRIS Data Management Center" developed for IRIS by the Science Horizons Corporation (Minster and Goff, 1986) and the TASC report (TASC, 1987) identified several guiding principles for a successful DMC.

Initially the concept of a large, self-contained Data Management Center was pursued with the understanding that,

- the task before it was formidable
- the budget for such a system would be greater than $10,000,000 per annum
- existing technologies within the reach of the university community could not manage the envisioned amount of data.

Now, nearly 20 years after the formation of IRIS, most of the original goals of data management within IRIS have been met or exceeded. Data volumes exceed the earlier projections by more than an order of magnitude, use of the system as measured by individual requests for data exceed expectations by more than two orders of magnitude and data from hundreds of recording systems are available in seconds to a few tens of minutes after real time. This was accomplished for a variety of reasons, not least among them that the seismological community retained tight control of the overall direction of the Data Management System and yet allowed a professional staff to take advantage of technological advances, achieving greater efficiencies than were imagined.

The structure of data management within IRIS has changed from the original centralized system that was envisioned to a hybrid system that takes advantage of both centralized and distributed components. While the IRIS DMC is still the largest component of the IRIS Data Management System, roughly one third of the financial assets of the IRIS DMS are provided to facilities outside the DMC. In the case of the permanent data from the Global Seismic Network (GSN) program of IRIS, two Data

Collection Centers (DCCs) are co-located with the Network Operations facilities in San Diego and in Albuquerque. This allows technical staff familiar with the details of the recording systems and their installation to be readily accessible to the technicians dealing with data and metadata issues. The three centers (IRIS DMC, ASL DCC and IDA DCC) form the heart of the IRIS DMS. The capabilities of these three centers are augmented via smaller and carefully monitored activities at US universities and in some cases, international data centers. Data quality assurance for data generated by the portable deployments of seismometers of the Program for Array Seismic Studies (PASSCAL) is funded directly by the IRIS PASSCAL program but strong and effective interfaces (people and computers) have been forged between the DMS and the PASSCAL programs.

The budget of the IRIS DMS was $3.4 million in fiscal year 2003. The IRIS DMS also receives $93 thousand per year to support activities within the Information Technology Research program of National Science Foundation (SCEC-ITR, 2001). We anticipate that the USArray component of Earthscope will also provide several hundred thousand dollars toward the activities at the IRIS DMC and the ASL DCC.

Historically the DMC has been able to augment its base funding through other sources. Notable among these include a Major Research Infrastructure (MRI) proposal to NSF for the replacement of the DMC mass storage system to a StorageTek Wolfcreek library in 1997, a Keck Foundation grant for $100,000 in 1997 and equipment and service donations from SUN Microsystems and StorageTek were used as cost matching for the MRI grant responsible for the purchase of the StorageTek Wolfcreek system.

At the beginning of 2003, the staff of the IRIS DMC numbers 15, the ASL DCC has 8 staff and the IDA DCC has 3.5 staff. Staff at the IRIS DMC and IDA DCC is fully funded from annual support from the NSF to IRIS. Financial resources from the United States Geological Survey (USGS) are used to pay for the staff at the ASL DCC but most major equipment used in the data collection activities at the ASL DCC are funded through IRIS.

DMC staff is divided into three primary groups. The operations group consists of four people who are responsible for archiving data and servicing requests for data from the user community. The software engineering group consists of seven people whose responsibilities include the development and maintenance of all software used within the

routine operations of the DMC, development of new user access tools, and development of new methods of serving data to the research community. The software group possesses strong computing skills that include relational database management systems, object-oriented software development, and CORBA distributed computing techniques. The final group of four people includes the DMS Program Manager, the Web-master, the DMC Office Manager and the UNIX Systems administrator.

Figure 2. The Powderhorn Robot has room for 6000 tape cartridges with capacities up to 200 gigabytes each. As such the total capacity of this system can be more than 1 petabyte (1,000,000,000,000,000 bytes).

A Brief History of Mass Storage at the DMC

In 1988 an Interim DMC was established at the University of Texas, Austin. While at this center, the preliminary techniques for managing the data from the GSN were developed. While in Austin, the DMC used the mass storage capabilities at the Center for High Performance Computing. The system developed around SUN Microsystems servers and SUN workstations. More than 15 years later, the IRIS DMC, and the DCCs are still primarily SUN-based. In 1991 the DMC acquired its first mass storage system. A Metrum RSS-600 running AMASS software was capable of storing 6 terabytes of information. This system served the DMC very well for nearly 5 years. Unfortunately it was the primary storage system for 6 years. The technology required to read the media became nearly impossible to maintain. The DMC learned the importance of insuring that data are routinely transcribed to newer technology storage systems roughly every 4 years, which is consistent with practice at other major data centers such as NCAR. It is not the life of the media that proved important, it is the ability to support the recording devices that truly controls the viability of an archiving system. In 1997 the IRIS DMC acquired a StorageTek Wolfcreek robot with helical scan Redwood tape drives and capable of storing 50 terabytes of data. In 2001 the DMC upgraded its storage robot to a 6000-slot capable Powderhorn robot with T9940 tape drives. This system was capable of storing 360 terabytes of data. As the technology in tape drives evolved

the DMC began transcribing data to higher capacity 9940B tape technology in 2003 and the robot's capacity grew to more than 1 petabyte (1 x 10^{15} bytes).

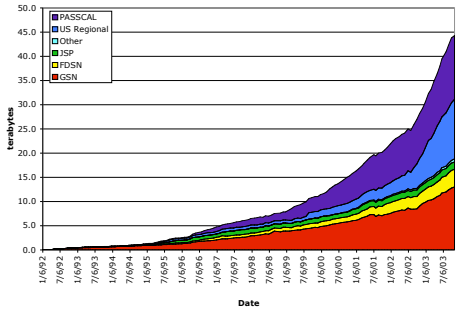

Figure 3. The archive of station and time-sorted data has grown exponentially with time. As of October 2003 there were nearly 45 terabytes of data in this archive. The GSN (red) and PASSCAL (purple) components form the heart of the archive. The FDSN (yellow), International networks (green) and US regional networks (blue) components provide roughly one-third of the data available as well.

The DMC data holdings in 2003 came from primarily five different sources. The IRIS GSN data holdings total 13.0 terabytes, the IRIS PASSCAL program holdings total 13.2 terabytes, regional networks within the US total 12.3 terabytes, networks from the FDSN have contributed 3.7 terabytes and other data sources have contributed roughly 2.1 terabytes to the DMC archive. As of November 2003, the archive contained almost 45 terabytes of data. The archive is stored in two sort orders, once by time and once by station that allows user requests to be serviced with high efficiency depending on the nature of the request. Each of the time and station sort orders are stored twice in the Powderhorn and the time sorted data are also stored on DLT tape in a secondary library. These DLT copies are transferred routinely to UNAVCO in Colorado for out-of-state safekeeping of all data holdings. The IRIS DMC currently manages more than 97 terabytes of data when the various copies and dual sort order of the data are taken into consideration.

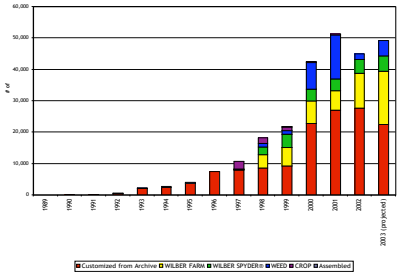

Figure 4. The shipments of customized products from the IRS DMC continue to increase. The slight decrease in 2002 was due to the absence of the WEED request mechanism in 2002.

The use of the IRIS DMC is very active, and it is considered to be the place most researchers go to obtain the data necessary to perform their seismological research. The original estimates believed that only a few hundred requests per year would be serviced by the DMC. As the figure above shows, the DMC is now shipping between 40,000 and 50,000 data shipments per year.