US ARRAY

Data Management Plan

Tim Ahern
Program Manager
IRIS Data Management System

IRIS DMC
1408 NE 45th Street
Suite 201
Seattle, WA 98105

August 2, 2004

Table of Contents

List of Figures

List of Tables

Executive Summary

The IRIS Data Management System (DMS) has 16 years experience managing seismological data. USArray data management will draw upon this experience to efficiently manage the many terabytes of seismic waveforms that will be generated annually by USArray.

USArray data will come from three primary sources. Data from the Transportable Array (TA) will flow from the field stations, normally in near real-time, from the ANSS Backbone (BB) or Permanent Network always in real-time, and from the Flexible Array (FA) through telemetry when available, or alternatively through tape transfer.

All USArray data will undergo quality control, sometimes in multiple stages and multiple locations, across various time scales. All USArray waveforms will ultimately end up at the IRIS Data Management Center (DMC), and user access to the waveforms will be through the IRIS DMC. Various levels of products will be produced from both the raw and processed waveforms, to be made ready for distribution.

USArray will follow the general procedures for data ingestion and distribution that the IRIS DMS has been following for many years. The major difference in handling USArray data versus many other kinds of data managed by the DMC is that the majority of data will be received and distributed in real-time in addition to delayed access through the archive. While other data sources are received in real-time, USArray data will be the first to have QA applied in near real-time.

All USArray seismic data going to or distributed from the DMC will conform to the International Standard for the Exchange of Earthquake Data (SEED) format, currently v2.4. SEED is a format that has been developed by the Federation of Digital Seismographic Networks (FDSN), a commission of IASPEI, a section of the IUGG, and is globally embraced.

All USArray data, regardless of the actual network that operates the seismic station, will be accessible through a concept called the Virtual Network, thereby simplifying user access to raw waveform data. Sub components of USArray data (TA, BB and FA) will all be available as virtual networks.

USArray Products will include raw waveforms, quality controlled waveforms, instrument corrected waveforms, event-segmented waveforms, earthquake locations and will evolve to include more complex products proposed by the USArray Community Products Working Group. This group will be composed of members of the seismological community, and will be coordinated through the IRIS DMSSC, the IRIS Coordinating Committee, and the IRIS Executive Committee. The IRIS DMC will be responsible for managing community defined products in a distributed product management system, as well as implementing a system that will be able to manage and distribute the full extent of these USArray products.

The USArray Data Management Plan

Earthscope is composed of three projects. The San Andreas Fault Observatory at Depth (SAFOD) is being developed by Stanford University and the US Geological Survey (USGS); the Plate Boundary Observatory (PBO) is being implemented by UNAVCO, Inc; and the United States Seismic Array (USArray) is being implemented by IRIS. This data management plan addresses the data management envisioned for the USArray component of Earthscope.

Strain meter and seismic data generated by other Earthscope components will also be managed by the IRIS DMC but this plan does not address this aspect of data management as completely as USArray generated data. Since the Earthscope MRE proposal specifies that the IRIS DMC will act as an archive for PBO strain meter data, that aspect is covered in greater detail than other data sets generated by PBO and SAFOD. Readers should refer to the PBO Data Plan and the SAFOD Data Plan for more detailed information regarding data generated by those projects. Nevertheless we do anticipate involvement of the DMC in managing SAFOD and PBO time series data.

USArray Components

USArray consists of four data generating components:

Permanent Network:

The ANSS Backbone is a joint effort between IRIS/USArray and the USGS to establish a permanent network of approximately 100 stations in the lower 48 states, plus additional stations in Alaska. The USArray contribution to the Backbone consists of 9 new GSN-quality stations, 4 cooperative stations from AFTAC and Southern Methodist University, and 26 stations from the Advanced National Seismic System (ANSS)

that will be upgraded with funding from Earthscope. The additional 60 stations of the ANSS backbone network either currently exist or will be installed or upgraded, and data will be made seamlessly available through the IRIS DMC. Data will be continuously recorded at 40 samples per second as well as 1 sample per second and transmitted in real-time back to the DMC. Quality assurance will be the responsibility of USGS -operated facilities at Albuquerque, New Mexico and Golden, Colorado.

Transportable Array:

The Transportable Array (TA) will consist of up to 400 stations, each station recording for an extended period of time, between 1.5 and 2 years. A total of approximately 2000 station locations will be occupied during the first ten years of USArray. Transportable Array stations will record 3 component seismic data at 40 samples per second and 1 sample per second continuously. Some of the TA stations will be from stations already currently in operation within existing regional networks. USArray will sometimes pay for upgrades to equipment or telemetry systems to assist in integrating these regional network stations. Data will normally flow from the stations to a Data Concentrating Node (DCN) over satellite or land-based circuits as appropriate in real-time. The DCNs might be local regional network operators, for example. From the DCN, data will flow into an Antelope Real-Time System and data will be simultaneously shipped to the Array Network Facility (ANF) at the Scripps Institution of Oceanography at the University of California San Diego, and to the IRIS DMC in Seattle. Quality assurance will be the responsibility of the ANF, including the function of assembling and managing the seismic station metadata needed by the DMC. The DMC will apply quality assurance techniques to all data received in real-time through its Quality Assurance Framework (QUACK).

Flexible Array:

The Flexible Array (FA) experiments will be conducted by individual PIs. Support for the experiments will be provided by the Array Operations Facility (AOF), operated in conjunction with the PASSCAL Instrument Center (PIC) at New Mexico Tech in Socorro, New Mexico. Numbers of instruments will vary. When possible, data will be telemetered in real-time to the DMC and simultaneously to the ANF. In other cases, data will be returned to the AOF via tape or other non-real-time methods. The AOF and ANF will be responsible for Quality Assurance of and metadata for the FA data.

Magnetotelluric Data:

Several magnetotelluric stations will also be co-located with seismic stations within the USArray. Data will flow from these sensors to the ANF where metadata, and quality assurance of the data, will take place before forwarding them to the DMC.

Additionally some backbone stations will be equipped with magnetotelluric stations and these data will receive quality assurance and managed by the IRIS DMC.

Data Policy

All USArray data will be open. Data from the Permanent Network and the Transportable Array will be available without any restrictions and normally in real-time. Data from the Flexible Array will allow the same 2-year proprietary period as data from PASSCAL experiments if specifically requested by the Principal Investigator (PI).

USArray Data Delivery Policy

The equipment in the USArray facility represents a significant community resource. The quality of the data collected by this resource is such that it will be of interest to investigators for many years. In order to encourage the use of the data by others and thereby make the facility of more value to the community, The Data Delivery Policy states that all data collected by these instruments should be available to the scientific community without undue delay.

Data from the Transportable Array (TA) will flow in "real-time" to the IRIS Data Management Center (DMC) and to the Array Network Facility (ANF). These data will be made available to the scientific community by the DMC as soon as they arrive. Data from the TA that cannot be transmitted by the real-time telemetry system will arrive at the ANF via disk or tape. These data will be processed by the ANF and sent to DMC as soon as possible after their arrival.

Data from the Backbone Array (BB) will flow to the DMC in near real-time and be made available without delay to the community.

Flexible Array (FA) deployments are different from the TA deployments in that they are PI driven and the PI is responsible for the operation and servicing of the experiment. In recognition of this, these data will be treated differently. All telemetered data from the FA will flow in "real-time" both to the DMC and ANF in the same manner as the TA data. The PI may request that the FA data be released only to the PI and his or her designees for a period of up to 2 years after the completion of the fieldwork.

Non-telemetered data from FA experiments will flow through the Array Operations Facility (AOF). The PI will make copies of all raw field tapes or disks and deliver these along with necessary field notes to the AOF at the completion of each service run. The AOF will process the data and deliver them to the DMC without delay. If the PI requests a delay, the DMC will withhold the data from public release for a period of up to 2 years from the completion of the fieldwork.

Data from major earthquakes of urgent scientific interest will be made publicly available through the DMC as soon after the event as possible.

Data Flow from US Array Components

As much as possible data will flow in real-time from the various USArray stations to the IRIS DMC. The following figure summarizes the seismic data paths involved in the entire USArray project:

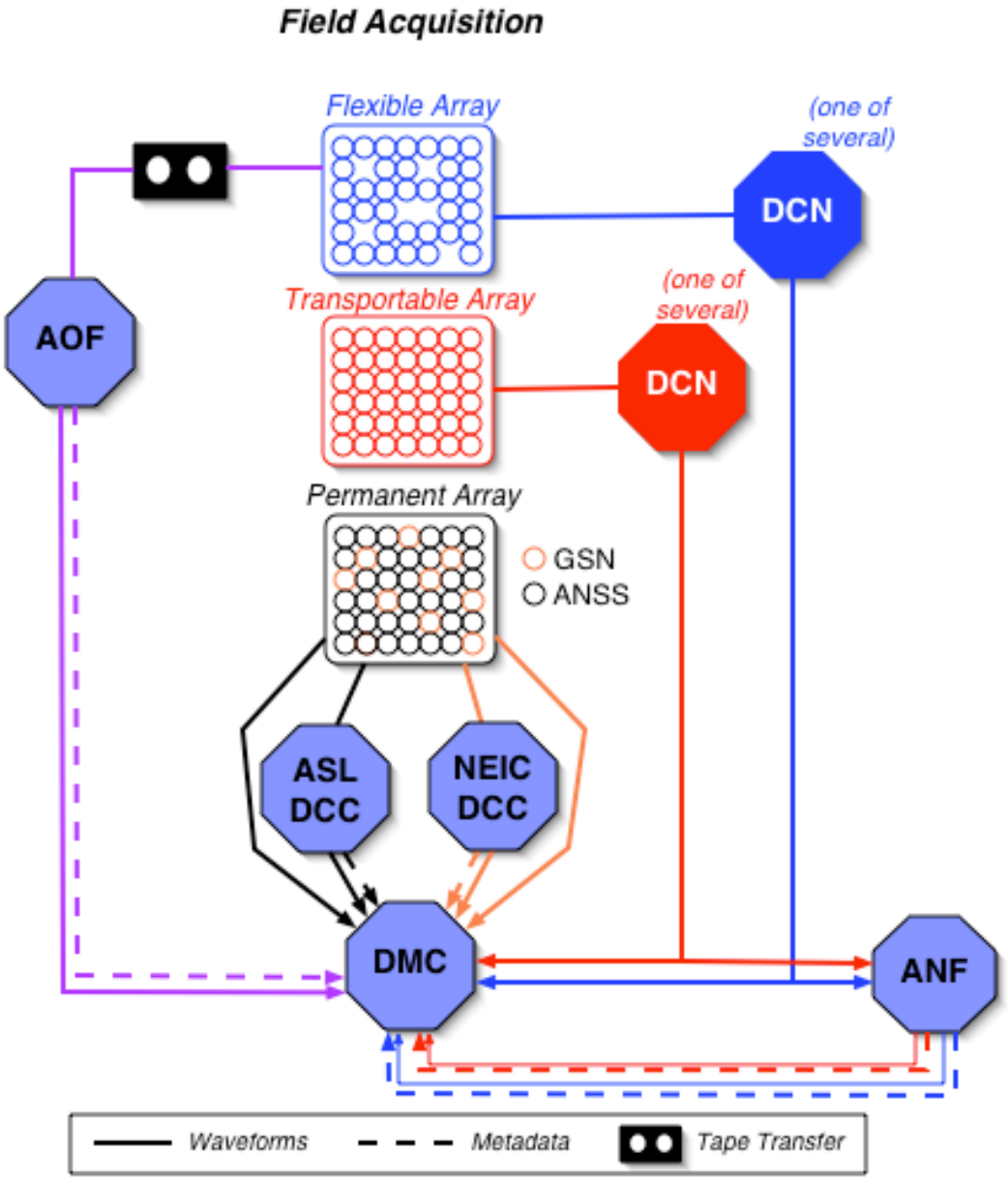


FIGURE 1. SEISMIC DATAFLOW WITHIN USARRAY.

This figure shows the flow of waveforms (solid lines) and metadata (dashed lines) within USArray. Data Concentrator Nodes (DCN) serve as real-time data communication repeaters. Data from the Transportable Array (TA) flow simultaneously to the ANF and the DMC.

All metadata for the TA come to the DMC from the ANF. Flexible Array (FA) data sometimes reach the DMC in real-time through a DCN but often will be received with some latency through the AOF. All metadata for real-time stations will flow through the ANF while metadata from the non real-time stations flows through the AOF. Data from the ANSS Backbone (BB) stations will flow in real-time to the DMC, NEIC DCC, and a subset to the ASL DCC. Quality controlled waveforms flow in a delayed manner from the ASL and NEIC DCCs to the DMC. All metadata comes to the DMC from either the ASL or NEIC DCCs.

ANSS Backbone Data Flow:

The BB consists of stations with either GSN quality or USNSN quality instrumentation. The primary difference in these installations is the sensor, and at times, the type of the installation. Data from the ANSS Backbone will flow from the stations over the ANSS satellite communications system to a downlink in Golden, Colorado. Data from 9 stations will be transmitted to ASL for data quality review. Data from 4 AFTAC stations will arrive at the DMC from AFTAC over existing dedicated data circuits, and then forwarded to ASL for quality assurance and review. Data from 26 USArray supported stations will be reviewed for data quality at the DCC in Albuquerque or in Golden. Data from the 60 additional stations within the ANSS backbone will be QA'd at the DCC in Golden.

Metadata will come from both the USGS ASL DCC and the USGS NEIC DCC, with each being responsible for only their respective stations.

Transportable Array Data Flow:

The TA consists of up to 400 stations at a given time. On average, one new station will be installed every day and one old station will be dismantled. Waveforms will flow from the TA stations over a communications circuit, usually a satellite circuit, to a Data Concentrator Node (DCN). At the DCNs, the ANF will develop the system to move data into an Antelope system, a commercial system developed by BRTT, Inc. Data will be simultaneously transmitted from the DCN to the IRIS DMC and the USArray ANF using Antelope ORB (Object Ring Buffer) to ORB technology. Some TA stations already belong to existing networks, so in those cases the DCNs may be collocated within a regional network operations center. However, data will still flow to the DMC and the ANF using Antelope ORB to ORB technology.

The ANF is responsible for all metadata for all TA stations. If the ANF modifies waveforms due to a repairable waveform problem, the ANF

will transmit the entire station day to the DMC using standard DMC procedures, irrespective of how many channels are affected. This will be the least common denominator for the cases where data need to be resent to the deep archive at the DMC.

Flexible Array Data Flow:

Flexible Array deployments are done by the PIs in a manner very analogous to existing PASSCAL experiments. As much as possible USArray will attempt to supply telemetry links to FA stations but it is anticipated that some stations will not be equipped with telemetry links. Telemetered FA data will flow to the ANF and to the DMC in parallel. Non-telemetered FA data will be sent to the Array Operations Facility (AOF) at the PASSCAL Instrument Center in Socorro, New Mexico. After QA the data will be transmitted to the IRIS DMC over standard Internet links. In general the DMC will only perform Quality Assurance on data being received in real-time into the DMC real-time BUD system. The ANF will support all FA real-time data.

Non-Seismic Data

Magnetotelluric Data

Some of the Transportable Array stations and Backbone stations may have co-located magnetotelluric installations as well. These data will flow to the ANF, where quality assurance tools developed by the Electro Magnetic Studies of the Continents (EMSOC) Consortium will be applied to the data, all appropriate metadata will be generated, and all data will be converted into SEED format. The data will then be forwarded to the DMC for archive and distribution.

Strain meter Data

The IRIS DMC will also act as a data archive and distribution point for strain meter data from the Plate Boundary Observatory (PBO). A duplicate copy of the strain meter data will also be archived at and distributed by the NCEDC in Berkeley. Two types of strain meter data will be generated as part of PBO. Borehole Strain meter (BSM) data will be converted to SEED format and transmitted to the DMC in real-time. Real-time borehole strain meter data will also go to a PBO facility in Socorro, New Mexico that will perform quality control on the data, and then send quality controlled versions of the data to the IRIS DMC. These duplicate data streams will be distinguishable using the Data Quality Control factor within the SEEDv2.4 format, where R implies raw data, and Q implies Quality Assured data. Laser Strain Meter (LSM) data will be converted to a PBO XML format and also sent to the

IRIS DMC for archiving. These data will reach the DMC in near real-time. A real-time stream of Laser Strain meter data will also be sent to a facility at UCSD where quality assurance will be applied and data will be sent to the DMC after quality control in the PBO XML format.

USArray Data Volumes

USArray will generate a large amount of seismic data each year of its existence. Table 1 provides a brief summary of the expected volumes of data that will be generated by USArray annually.

	Transportable	Permanent	Flexible	MT
Year 1	0.164	0.030	0.158	0.000
Year 2	0.569	0.119	0.979	0.003
Year 3	1.031	0.217	1.847	0.027
Year 4	2.077	0.235	3.867	0.061
Year 5	2.529	0.235	4.210	0.061
Year 6	2.529	0.235	4.210	0.061
Year 7	2.529	0.235	4.210	0.061
Year 8	2.529	0.235	4.210	0.061
Year 9	2.529	0.235	4.210	0.061
Year 10	2.529	0.235	4.210	0.061

Table 1. USArray Annual Data Generation in Terabytes.
The above table shows data volumes in terabytes from the USArray subcomponents. When fully deployed, USArray will generate roughly 7 terabytes of data per year from the various sub-components. All of these data will be archived and managed at the IRIS DMC in Seattle. Currently (2004) the IRIS DMC is ingesting about 8 terabytes of data per year. When USArray reaches its full compliment of stations it will only double the amount of data flowing into the DMC.

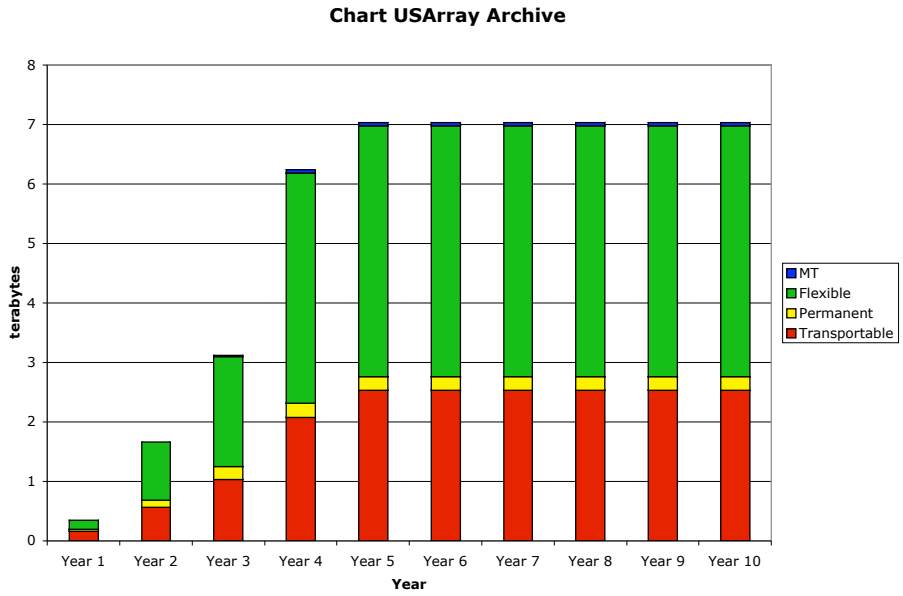


FIGURE 2. ANNUAL DATA GENERATION BY USARRAY IN TERABYTES

This graph presents the information found in Table 1. During the initial 4 years of USArray, stations are still being acquired and USArray is not at full strength until the 5th year. Again this only shows the single sorted values, the dual sorting employed in the DMC data management strategy doubles these volumes.

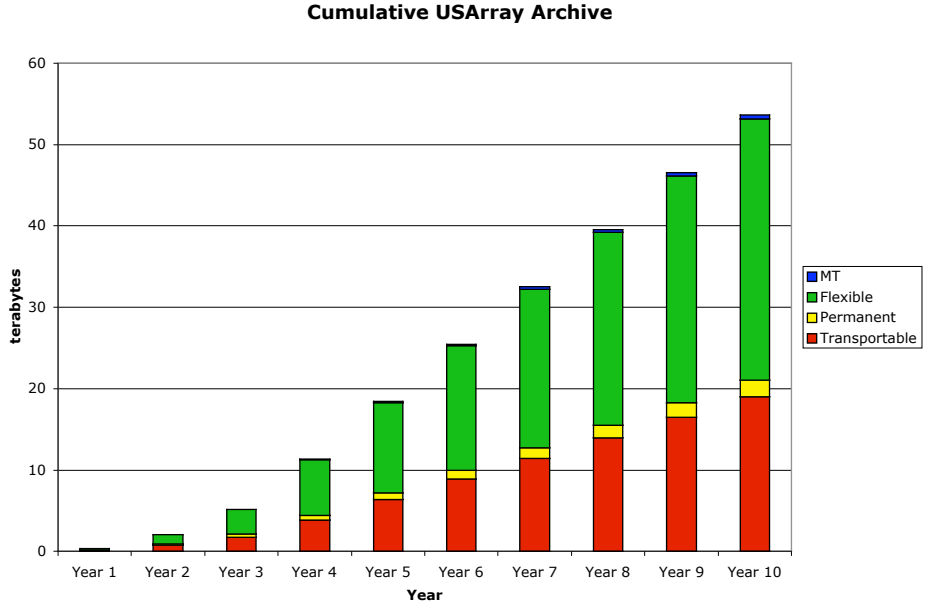


FIGURE 3. CUMULATIVE USARRAY DATA VOLUMES IN TERABYTES.

This graph shows the cumulative volume of USArray data over the first ten years of its existence. By year 10, a single copy of USArray data will require roughly 54 terabytes to store.

Data Quality Assurance

Procedures for assuring data quality are very mature within the IRIS DMS. USArray data will receive quality assurance at multiple locations and by multiple centers, depending on which specific element of USArray is generating the data. The specific type of QA varies at the various facilities but, in general, is NOT redundant.

Permanent Array:

Data from 13 of the USArray supported ANSS Backbone stations (9 GSN style and 4 AFTAC style) will be received in real-time. Additionally, data from 26 enhanced ANSS Backbone stations will also receive QA as they flow into the IRIS Data Management System. These streams will flow to the Albuquerque Seismic Laboratory operated by the USGS, where GSN style QA will be applied. This includes routine estimation of timing, sensor orientation, sensor calibration and general waveform quality issues. ASL will maintain the station metadata database for these stations. QA'd data will be forwarded to the DMC with roughly a 24 hour latency.

Data from 60 USGS-supported Backbone stations will flow in real-time to the NEIC Data Collection Center in Golden, CO. In Golden, similar QA will be applied to the data as will be applied in Albuquerque. Again, QA'd data will be forwarded to the IRIS DMC with about a 24 hour latency. All of these data will also be forwarded to the IRIS DMC for archiving and distribution. In addition to this QA, all of the raw data streams will flow in real-time to the IRIS DMC where the QUality Assurance Control Kit (QUACK) system will make routine QA measurements on the data. These measurements will be stored in tables within the primary DMC Oracle RDBMS and will be accessible to users of the DMC. Raw data will be available through the DMC as soon as technically reasonable (usually from 1 to a few minutes) from the IRIS DMC, going to all users that wish to receive it. The QA'd data will also be available through the DMC within 24 hours of real-time. The Data Quality Flag in the SEED format will be used to distinguish data that have received quality assurance procedures from those that have not.

Transportable Array:

Data from the TA stations will normally flow in real-time to a Data Concentrator Node (DCN) and simultaneously flow from the DCN to the IRIS DMC and the ANF using Antelope ORB to ORB technology. These real-time data will be converted into miniSEED format when they are extracted from the Antelope ORB. The ANF will perform quality control on all of the TA data including:

- routine phase picking of waveforms,
- timing of later arriving phases,
- hypocenter determination
- waveform quality assurance
- timing problems
- station specific characteristics

The ANF is responsible for all metadata from the TA stations.

Existing regional networks within the United States will be responsible for the operation of some of the stations of the TA. In these instances, the regional networks will in effect become a DCN from the perspective of USArray operations. Data will be transmitted simultaneously to the ANF and the DMC from these regional centers. The ANF will still be responsible for the metadata for the stations working in conjunction with the regional center, and will become the authoritative source for the DMC during the time period in which existing stations are included within the USArray network.

The raw data streams will flow in real-time to the IRIS DMC where the QUACK system will make routine QA measurements on the data. The use of the QUACK system for TA data is completely analogous to its use for BB data.

Flexible Array:

Data from the FA will sometimes be telemetered and sometimes not. The ANF will support real-time data for the FA as it does for the TA. The Array Network Facility (ANF) and the IRIS DMC will simultaneously receive telemetered data. The ANF will be responsible for QA and for all metadata for telemetered FA data. For non-telemetered data the AOF will received the data from the field, perform QA, and prepare the metadata. The DMC will not perform additional QA on these non-telemetered FA data, and will only act as the archiving and data distribution center.

Magnetotelluric Data:

Data from the MT stations will be QA'd within the ANF. The ANF will translate all data from native format into the SEED standard using systems previously developed by the EMSOC community.

Strain Meter Data

Borehole strain meter data will receive QA within the PBO supported center in Socorro. Both un-reviewed real-time and quality-controlled data will be archived at the DMC. Laser strain meter data will be QA'd at a center at UCSD and the QA'd data will be transmitted to the DMC for archiving and data distribution.

IRIS DMC Quality Control Framework

The IRIS DMC has developed a system called the Quality Assurance Control Kit (QUACK). At the present time it works only on data entering the DMC real-time system, the Buffer of Uniform Data (BUD). TA, BB, and a portion of the FA data will be received through the BUD system in real-time and as such will receive automated quality control. QUACK is a framework from which a variety of plug-in Quality Control applications may be invoked. QUACK controls the timing as to when QA applications are executed, is responsible for presenting the desired data to the QA Application and is responsible for storing measured parameters within the IRIS DMS Oracle RDBMS. More information about QUACK can be found at

http://www.iris.edu/QUACK/tutorial.html.

At the present time QUACK measures parameters such as RMS and mean values, data glitches, data dropouts, data overlaps, percent availability, power spectral density functions, timing quality, data continuity check, and a wide variety of additional tools that include some data format validation. The specific QA functions can be expanded and modified as needed, as QUACK is very flexible. The goal is to evolve toward an automatic system that can identify most problems affecting data quality and alert a data technician so the problem can be corrected or eliminated in a timely manner.

The output of QUACK is placed in the IRIS DMC Oracle RDBMS and reports, graphics and alerts can be generated from the RDBMS. We anticipate that QUACK will be used to alert Data Technicians of data problems, reducing the requirement for extensive data analyst oversight, and will trigger human intervention only when problems arise, and will facilitate more immediate problem resolution.

Data Archiving

Part of the DMC's responsibilities is to insure that the data are archived and protected in-perpetuity, and in a manner to insure that data are never lost. Elements of the DMC archiving strategy to enable this reliability include:

1) Offsite copies – copies of BB and TA data are retained at the ASL/NEIC DCCs and the ANF respectively. Copies of the strain meter data will also be held at the Northern California Earthquake Data Center (NCEDC). Copies of the MT data will be retained at the ANF.
2) IRIS DMC Sort Redundancy – all waveform data are stored in two sort orders within the IRIS DMC mass storage system, once by time and once by station. If needed, for example as a result of media failure, a version of one order could be created from a version of the other sort order. The primary reason for two sort orders, however, is to efficiently service requests for these data.
3) IRIS DMC Onsite Archive copies. The DMC stores a duplicate copy of each of the sort orders within the primary mass storage system.
4) IRIS DMC Offsite Backup. The IRIS DMC stores one copy of the time-sorted data on a different media. This backup copy is shipped once a week to the UNAVCO facility in Colorado.

An important criterion for offsite storage requires that all copies of data can be read with UNIX tar alone even if the primary mass storage control application becomes corrupt. It would be time consuming, but all data could be recovered.

The DMC also records data on at least two different types of tape media. Presently we store data on StorageTek 9940B tape media as well as Linear Tape Open (LTO) tape technology. Therefore, a problem with one tape technology can never cause a serious problem with DMC archive holdings.

The DMC also stores copies of key software application source code, scripts, and executables within the mass storage system. Copies of key programs and database files are also stored on backup media and transferred to commercial data safes, located in Eastern Washington, on a weekly basis.

Data Transcription:

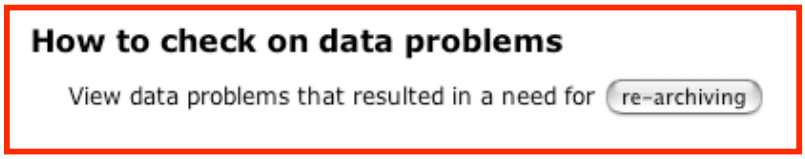
The DMC archiving plan includes transcribing data at least every 5 years (normally 4 years) to new media, and when necessary, to new mass storage recording technology in order to leverage state of the art technologies and to guarantee both reliability and perpetuity. We have just begun our fifth data transcription in the 16 years the DMC has existed.

As a result of the careful and redundant archiving and storage policies, no data have been lost once it has been archived at the DMC during its 16 years of existence.

Data Replacement

When problems with waveform data are encountered at one of the various QA nodes within the USArray system, replacement data will be electronically transmitted to the IRIS DMC using existing procedures in place at the DMC. Whenever waveforms are to be replaced, they will be replaced on a station-day granularity, meaning ***all*** data a station generated on the day in question will be resent to the IRIS DMC.

The IRIS DMC will document the re-archiving activity on the existing re-archiving section of the IRIS web site. Access this page http://www.iris.edu/data/data.htm and select the re-archiving option.



Data Access Systems

Real-time

The IRIS DMC operates the Buffer of Uniform Data (BUD) real-time system. This system will be leveraged heavily within USArray. The following diagram shows waveform data flow from the various USArray components into the USArray BUD system. We anticipate roughly 7 terabytes per year of data flowing directly into the BUD system. Real-time data will reside in the BUD for roughly 30 days, but will be moved to the archive system 6 days behind real-time.

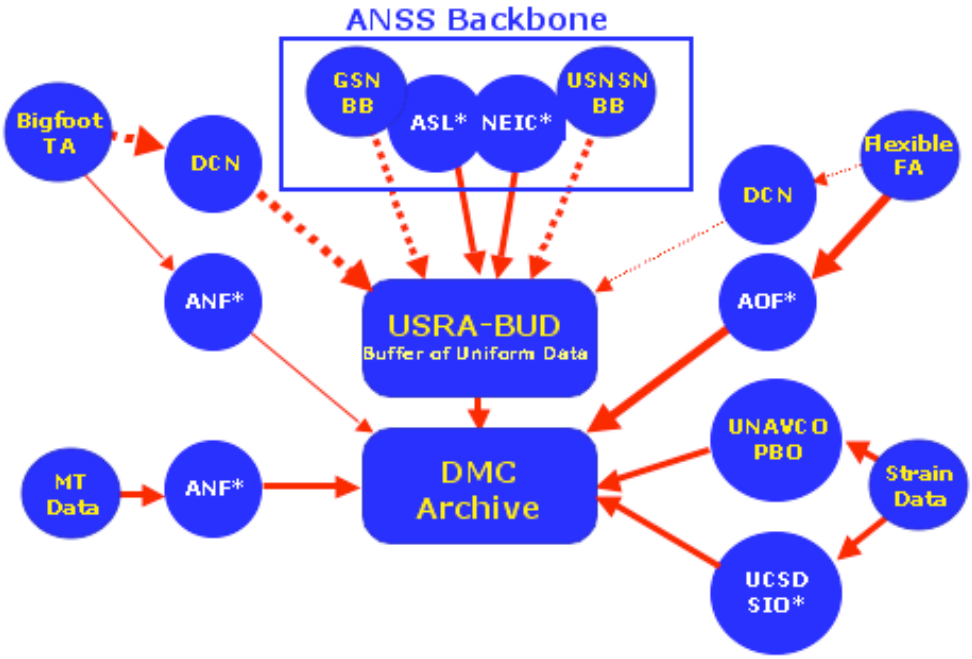


FIGURE 4. DATA FLOW INTO THE USARRAY BUD AND ARCHIVE SYSTEMS.
USArray data will flow into the IRIS DMC along multiple paths. Backbone data will flow in real-time to the DMC as well as through the ASL DCC or NEIC DCC, receive quality control, and be forwarded to the DMC. Transportable array data will be simultaneously telemetered in real-time to the DMC and ANF from a Data Concentrator Node (DCN). Transportable Array data that are not telemetered will travel to the ANF, be processed as necessary and sent on to the DMC. Flexible array data will mostly be sent through the AOF and a smaller proportion of the data will be received in real-time at the DMC and ANF. The DMC will also receive data from the PBO strain meters from UNAVCO and UCSD, and MT data from the ANF/EMSOC facility. Facilities that are responsible for providing and validating the metadata are indicated with * in the above diagram. Dashed lines indicate real-time telemetry paths and solid lines indicate delayed transfer. The width of the lines give an indication of dominant data path for each source, either real-time telemetry or delayed transmission.

Archive Access

The IRIS DMC has a rich variety of data access tools including those that are email-based, web-based as well as through programmatic interfaces. All existing access tools will be able to service requests for USArray data and PBO strain meter data. We do not anticipate adding new tools to specifically allow access to USArray data but rather leverage the existing data access infrastructure at the DMC. Figure 5 displays the variety of tools currently available.

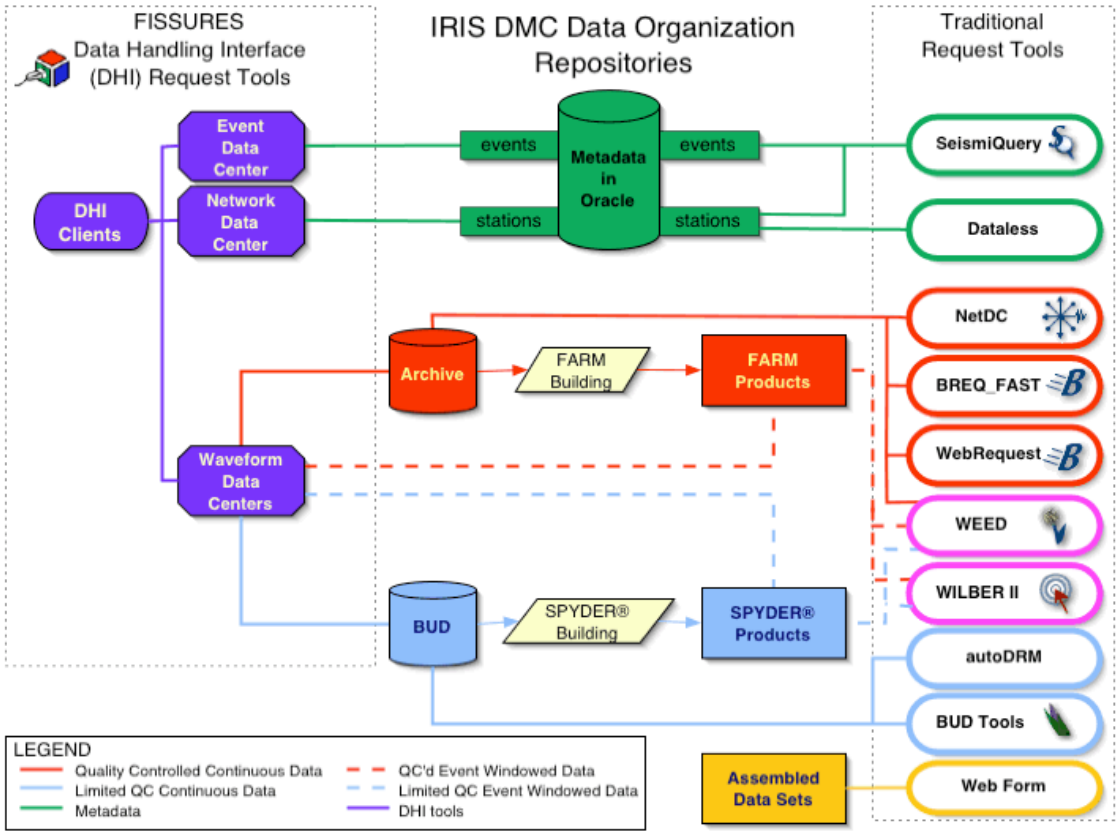


FIGURE 5. IRIS DMC DATA ACCESS TOOLS.

This figure shows the powerful variety of access tools, through which the scientific community will be able to gain access to data from USArray. This figure is available at

http://www.iris.edu/data/req_methods.htm

and is clickable from that location. A complete Data Access Tutorial is available on the web at

http://www.iris.edu/manuals/DATutorial.htm

and provides a comprehensive guide to using these tools. The lines that connect the request tools (on the extreme right and left of the diagram) to the repositories indicate which tool can be used to access the various repositories.

As the real-time demand for data from substantial portions of USArray increases, we will be developing a DHI-based real-time data distribution system, much of which is already working. We will also investigate exploiting various Internet telemetry protocols, such as multicasting, as a mechanism for distribution of both real-time data and USArray products.

Virtual Seismic Networks

USArray is composed of three primary sub-networks, the BB, TA and FA. The Backbone itself is well coordinated with AFTAC, the USGS and regional networks within the United States since these organizations

actually own and operate stations. To maintain the correct attribution for the data, the network code of the waveform files identifies the operator of the network and will not necessarily indicate that the station is part of the USArray Network.

For these reasons the IRIS DMS is developing the concept of Virtual Seismic Networks. Through the virtual seismic networks, the various sub-networks of USArray can be defined as containing stations that are actually part of another network. For instance, the initial BB network initially consists of 4 stations that are operated by AFTAC (network code IM) and 4 stations that are operated as part of the USGS ANSS backbone (network code US). Without the successful implementation of the Virtual Seismic Network concept, it would quickly become difficult for a researcher to easily access all the data from a USArray sub-component, or for that matter, all of the data from USArray.

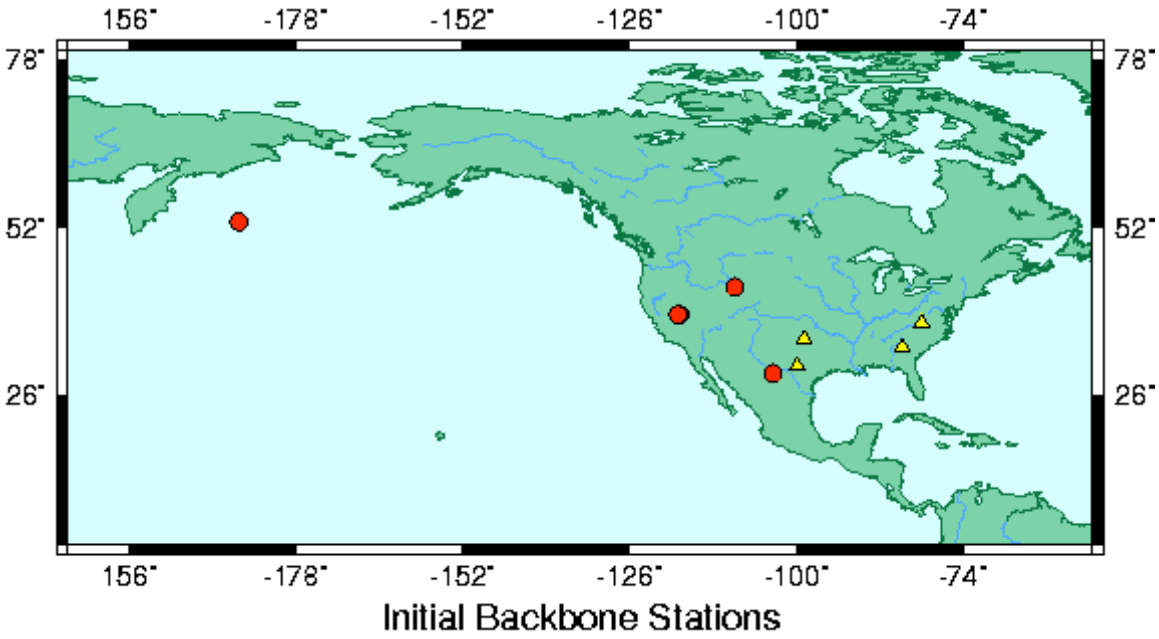


FIGURE 6. THE INITIAL BB VIRTUAL NETWORK.

The initial backbone network will consist of 4 stations from AFTAC (red circles) and 4 stations from the ANSS (yellow triangles). While these stations have different Network Codes (IM and US respectively), users of the data will be able to specify _US-BB as the Virtual Network identifier and access all data from the USArray backbone.

All stations that belong to the USArray Backbone will be accessible by specifying the virtual network code *_US-BB where one would traditionally specify a 2 character Network Code such as US, CI, or BK.*

The first stations of the TA will be stations operated by Caltech (network code CI), UCSD (network code AZ) and Berkeley (network code BK). As the TA moves across the country, other stations from existing networks will also be incorporated within the transportable component of USArray.

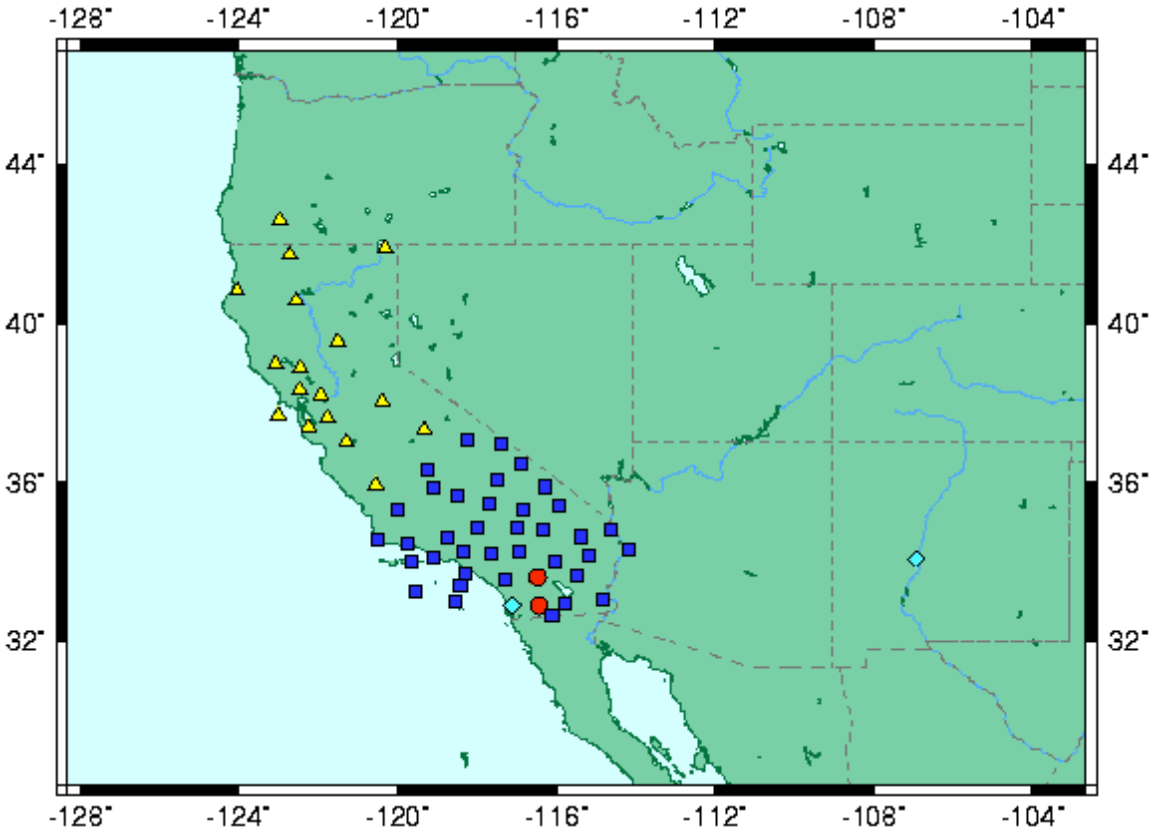


FIGURE 7. TRANSPORTABLE ARRAY IN CALIFORNIA

This figure shows the locations of the initial transportable array stations in California. In addition to two new stations (light blue), the Transportable Array also has several stations contributed by Caltech (dark blue), Berkeley (yellow), and UCSD (red). (From BUD System July 23, 2004)

All TA stations will be available by specifying _US-TA as the network code. The following table shows some of the existing Virtual Networks that have been defined.

Virtual Network Name	Network Description
_US-TA	All components of the Transportable Array including Regional Networks and new TA stations
_US-BB	All stations in the Backbone Array including some regional network stations, some ANSS stations and designated AFTAC and GSN stations
_US-FA	All stations that are part of the Flexible Array
_US-ALL	The union of _US-TA, _US-BB and _US-FA. This would give a data requester data from all stations that are connected with USArray in any way.

Table 2. Virtual Network Definitions.

This table shows Virtual Networks that have been defined for USArray. The purpose of Virtual Networks is to leave the time series labeled with the network code of true seismic network operator for attribution and correctness but still allow easy ways for data requesters to gain access to all stations that belong to USArray components, without having to maintain complicated lists of stations. While this table shows all networks related to USArray that are currently defined, we anticipate that virtual network definitions will continue to be added.

Metadata

While the majority of data in the IRIS DMC are seismological time series, the more difficult to develop and maintain portion of the data center is the metadata that describes the waveforms and the stations themselves.

IRIS is a major participant within the ICSU/IUGG/IASPEI commission called the Federation of Digital Seismographic Networks (FDSN) (http://www.fdsn.org). Among the many functions that the FDSN plays is to coordinate data exchange formats across the international community. A Working Group within the FDSN that has IRIS and USGS representation defines the resulting SEED format. This format is also being adopted for use within the Advanced National Seismic System (ANSS) of the USGS.

One of the key items that have resulted from the adoption of the SEED format is that the majority of the necessary metadata that are needed to describe the data are provided within the SEED format.

USArray has adopted the SEED format as the waveform format within USArray, and all waveform data will be distributed in that format. Format conversion utilities already exist to convert from the SEED exchange format to the most common analysis formats.

The various QA nodes within the dataflow path (Figure 4) are responsible for generating and maintaining the metadata required in the SEED format. The IRIS DMC has a mature system for ingesting these data into the data management system, and specifically into the Oracle DBMS at the heart of the DMC.

We will likely support XML-SEED that is being developed by the Japanese, and as such, the rich metadata will be available in XML as well as traditional SEED.

The IRIS DMC generates a large amount of metadata during the data ingestion process, the most significant being indications of the starting time and ending time of continuous segments of waveforms from the thousands of channels that are managed. The QUACK framework also makes a large variety of measurements on time series from the various stations whose data are at the DMC. All of these metadata that relate to data quality are available from the IRIS DMC systems. As part of the need for the IRIS DMC and ANF to develop and manage additional station metadata, the two nodes are working together to develop the IRIS/USArray Station Information System (ISIS). This application will be able to track metadata that is station specific and not currently capable of being contained in the SEED metadata. While some information within ISIS will be restricted (contact information of landowners for instance), most of the important information will be available over the Internet. ISIS maintains data locally in a relational database but can export the information in XML.

Data Products

As part of the MRE EarthScope award, the IRIS DMS will produce a limited number of data products, usually automatically. USArray Products are categorized in the following manner:

Level 0 – Raw Waveforms

The lowest level products of USArray are seismic time series or waveforms. In general these will flow directly from the stations to the DMC in near real-time and be managed in either the BUD repository or the SPYDER® repository as shown in Figure 8.

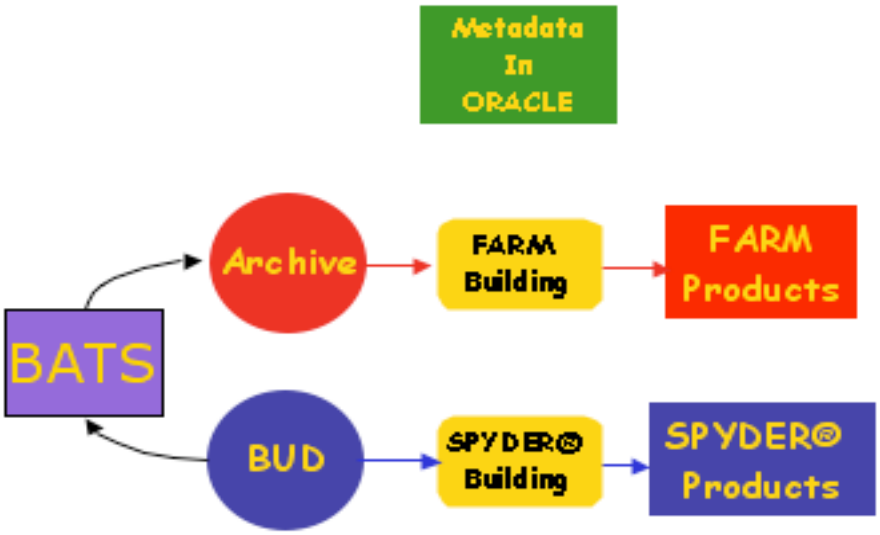


FIGURE 8. WAVEFORM REPOSITORIES AT THE IRIS DMC.
This diagram shows the four waveform repositories at the IRIS DMC into which USArray data will flow. Continuous real-time data will flow into the IRIS BUD (Buffer of Uniform Data). After QA and a delay of 7 (seven) days, the continuous data will move into the 1-petabyte mass storage system at the DMC. Whenever events of significant magnitude occurs, event alert messages received from the National Earthquake Information Center (NEIC) are sent to the IRIS DMC where the SPYDER® system works to trigger and segment event waveform files automatically. Roughly five weeks behind real-time, the NEIC transmits the weekly Preliminary Determination of Epicenter (PDE) earthquake catalog to the DMC. This catalog triggers the FARM (Fast Archive Recovery Method) product generation. The data from the FARM is obtained from the Archive and in general only contains data that have been quality controlled. All Metadata are managed in an Oracle Database Management System, and appropriate metadata are returned to the data requester whenever SPYDER® or FARM products are requested.

Level 0.0 Raw Continuous Data (BUD)

All continuous waveforms will be available from the DMC within a few seconds or minutes after real-time, depending on the

particular data transmission systems that can impose network (or packet) latencies. Waveforms will be tagged with a SEED Data Quality Factor of "R" to designate they are raw waveforms. These data will be available within the IRIS DMC BUD system with latencies of a few seconds to a few minutes after real-time, and will remain for roughly one month. This allows for uninterrupted, online access, as well as sufficient time for backfilling as a result of non-heterogeneous latencies. After six days, a copy of the data will be transferred to the IRIS DMC Archive after having QUACK Quality Assurance applied to them. Access to the continuous data will be available in perpetuity from the IRIS DMC archive.

Level 0.1 Raw Event Gathers (SPYDER®)

When events of significant magnitude occur, the IRIS SPYDER® system will be invoked for USArray Data. The magnitude threshold is currently set to all events above 5.5 worldwide, but a lower threshold (yet to be determined) will be set for earthquakes within the United States. Whenever an event with a magnitude exceeding the USArray criteria occurs, all raw waveform data for the event will be assembled into a SPYDER® waveform event volume. Access to these data will be provided by the WILBER, WEED and DHI data access methods, a description of which is available online at

http://www.iris.edu/data/req_methods.htm .

Level 0.2 Raw Data Quality Assurance Estimates (QUACK)

As time series data flow into the BUD real-time data system, they will be passed through a variety of quality assurance tools. There are currently 15 different algorithms that are run on all USArray data. In general these QA tools measure specific aspects of data quality that can be represented by a small set of numbers. These values are stored in an Oracle DBMS and will be made available to the scientific community. These estimates can be considered to be derived metadata that can be used to mine the USArray data set in a variety of ways.

Level 1 - Quality Assured Waveforms

All USArray data will undergo quality assurance procedures. QA'd data from the Backbone will be available from either the ASL DCC or the NEIC DCC roughly 24 hours after real-time. The Transportable Array data will be available with ANF and DMC QA applied roughly 6 days after real-time from the archive repository shown in Figure 6. Flexible Array data may experience significant delays before being available, as

a result of collection and re-formatting. Once in the archive, all USArray data will have undergone quality control.

Level 1.0 Quality Controlled Continuous Data (Archive)
The IRIS DMC will have continuous quality controlled data from all USArray components available to the scientific community with the latencies specified above. All request tools that access the archive (see figure 5) will enable access to the quality-controlled data in the archive.

Level 1.1 Event Gathered Data (FARM)
In a manner similar to SPYDER®, FARM event segmented volumes are extracted from the quality controlled archive whenever events of significant magnitude are encountered. These volumes are usually produced 5 weeks behind real-time when the NEIC distributes its Weekly PDE catalog, and allows for as much data as possible to become available. These products are managed so that late arriving or updated data are always added to the SPYDER® and FARM products when available.

Level 2 – Products Derived from Waveforms

Level 2 products are those that are derived from level 0 or level 1 products by procedures that are in general non-controversial and well understood. These procedures will be applied within nodes of the IRIS DMS.

Level 2.0 Instrument Corrected Waveforms
Level 0 and 1 products are in digital counts and have not been corrected for the response of the various filters in the seismometers and acquisition systems. Additionally the FIR filters used in modern digitizing systems introduce acausal effects. For this reason an application will be developed that can correct time series for the instrument response. This tool will allow users to gain access to any data in any of the four DMC repositories and correct them for the instrument response. Offering this as a Web Service is one concept that will be investigated.

Level 2.1 Instrument Corrected Event Segmented Waveforms
In a manner analogous to SPYDER® and FARM products, windowed data will be extracted from the Level 2.0 Instrument Corrected Waveforms and made available as products.

Level 2.2 Record Sections
The ability to generate customized record sections for all SPYDER®, FARM and Instrument Corrected event segmented waveform product will exist. Users will be able to select different types of filters to apply to the waveforms.

Level 2.3 Event and Station Related Products
The IRIS DMS will produce a variety of products that are routinely produced for seismic events. These will include things such as

- Phase picks made from automated systems such as Antelope
- Hypocenter determinations
- Centroid Moment Tensor solutions

Other types of information that will be determined include:

- Site characteristics at USArray stations
  - By locating many events with wide variability in their azimuth from a station, the ANF will statistically assess site-specific geological variations for each USArray station.
- Ground Noise at USArray stations
  - Routine noise estimations at USArray stations will be made by the ANF and by the IRIS DMC Quack system.
- Station Instrumentation
  - Channel calibration will continuously be monitored by the ANF and changes reflected in the metadata they provide to the DMC.
  - Sensor orientation will be determined by correlating energy arriving from events with known hypocenters, and corrected if necessary.

The IRIS DMC, the ANF or other components of the IRIS DMS will produce levels 0, 1 and 2.

Levels 3 and 4. Community Defined Products

USArray will produce an unprecedented collection of observational data that will result in improved understanding of the North American continent and margins, crossing many geological and geophysical domains. We anticipate wide-ranging participation in the development

of data products beyond the levels identified above. The IRIS Executive Community has tasked the Data Management System Standing Committee (DMSSC) to organize a working group that will take the steps necessary to identify the community defined products that will come from USArray. This committee was only recently populated and a workshop will be held in the second half of 2004 to define these products.

Level 3 products will be those that are Interpretive Products that require technical analysis and interpretation of the USArray data to be produced.

Level 4 products are Knowledge Products that will most likely require more integrative, cross-disciplinary effort in order to make the information meaningful to a very broad range of end users.

It is anticipated that some of the products will be generated at a centralized location such as the IRIS DMC or the ANF, while others may be generated at a scientist's home university, for example. From the outset, we are anticipating the need to develop a distributed environment within which these heterogeneous products will be generated and distributed.

Uniform Product Distribution System

As part of the developments within the IRIS DMS we will develop the Uniform Product Distribution System (UPDS). This system will be a fairly complete web service implementation including leveraging technologies such as XML, SOAP, WSDL and we hope an instantiation of UDDI. The UDDI will act as a yellow page directory from which individuals or applications can discover resources such as USArray products on the Web, determine how to use them, and even manipulate them through other Web Services.

Glossary of Terms

9940B	A 200gb/tape tape format supported by Storage Tek
AFTAC	Air Force Technical Applications Center
ANF	Array Network Facility at UCSD in San Diego, CA
ANSS	Advanced National Seismic System of the USGS
Antelope	A seismic network software system developed by BRTT (ARTS)
AOF	Array Operations Facility at New Mexico Tech in Socorro, NM
ASL	Albuquerque Seismological Laboratory (USGS)
BATS	BUD to Archive Transfer System
BB	BackBone Array, a sub-network of USArray and part of the permanent station in the ANSS Backbone
BRTT	Boulder Real-time Technologies, developer of Antelope
BSM	Borehole Strain meter
BUD	Buffer of Uniform Data, the IRIS Real-time System
DBMS	Data Base Management System, such as Oracle
DCC	Data Collection Center
DCN	Data Concentrating Node
DHI	Data Handling Interface, a CORBA based access tool at IRIS DMC, and elsewhere.
DMC	IRIS Data Management Center in Seattle, WA
DMS	Data Management System (IRIS)
DMSSC	DMS Standing Committee
EarthScope	The Major Research Equipment Program funded by NSF
EMSOC	Electro Magnetic Studies of the Continents
FA	Flexible Array – sub-network of USArray, PI driven
FARM	Fast Archive Recovery Method, event segmented data repository
FDSN	Federation of Digital Broadband Seismographic Networks
GSN	Global Seismographic Network (IRIS)
IASPEI	International Association of Seismology and Physics of the Earth's Interior
ICSU	International Council of Scientific Unions
IGPP	Institute of Geophysics and Planetary Physics
IRIS	Incorporated Research Institutions for Seismology, operator of USArray
ISIS	IRIS Station Information System
IUGG	International Union of Geodesy and Geophysics
LSM	Laser Strain Meter
LTO	Linear Tape Open, currently a 200 Gigabyte tape technology
MRE	Major Research Equipment, the Large Facilities Program with NSF
NCEDC	Northern California Earthquake Data Center at UC Berkeley
NEIC	National Earthquake Information Center (USGS)

ORB	Object Ring Buffer, a part of Antelope
PASSCAL	Program for Array Seismic Studies of the Continental Lithosphere (IRIS)
PBO	Plate Boundary Observatory- geodetic component of Earthscope
PI	Principal Investigator
PIC	PASSCAL Instrument Center at New Mexico Tech in Socorro, NM
PSD	Power Spectral Density
QA	Quality Assurance
QUACK	IRIS DMS' Quality Assurance Control Kit
RMS	Root Mean Square, a statistic
SAFOD	San Andreas Fault Observatory at Depth, drilling component of Earthscope
SEED	Standard for the Exchange of Earthquake Data
SIO	Scripps Institution of Oceanography, at UCSD
SOAP	Simple Object Access Protocol
SPYDER®	System to Provide You Data from Earthquakes Rapidly, an event oriented data repository at IRIS
TA	Transportable Array – sub-network of USArray
UCSD	University of California at San Diego
UDDI	Universal Description, Discovery and Integration
UNAVCO	University Navigation Consortium, Operator of PBO
UPDS	Uniform Product Distribution System
USGS	United States Geological Survey
USNSN	US National Seismic Network, predecessor to the ANSS
WSDL	Web Services Description Language
XML	eXtensible Markup Language