1.4. Data Management

Among the fundamental principles underlying the collection and distribution of EarthScope/USArray data are that:

- All data will be freely and openly available to all interested parties.
- Wherever possible, data will be collected and distributed in near-real-time.
- Data (including both time series from sensors and associated metadata) will be reviewed to ensure quality.
- All data will be archived at the IRIS Data Management Center (DMC).
- All data will be distributed by the IRIS DMC using both traditional and newer Data Handling Interface (DHI)-based data distribution methods.

In the same way that many of the technical standards underlying the USArray field systems have emerged from the PASSCAL and GSN programs, the collection, distribution, and archiving of USArray data will be based on procedures developed by the IRIS Data Management System (DMS). Hardware and software systems now in use for handling PASSCAL and GSN data will be augmented and expanded to incorporate USArray data. In addition to leveraging investments already made by NSF in developing the DMS, this will also ensure that users will be able to merge USArray data with existing data resources in a simple manner, and access to USArray data will be via familiar and well-tested procedures and tools

Data Volumes

Table II-1.1 summarizes the anticipated data volumes from the three components of USArray. The aggregate data flow rate is estimated to 4.2 terabytes per year. The IRIS DMC currently archives approximately 3.5 terabytes per year of seismic waveform data so that the total output from the fully installed USArray will roughly double the DMC's current rate of data collection and archiving. Because of the modular hardware configuration and highly automated procedures established at the DMC, this increase in data flow can be incorporated with relatively minor increases in staffing and hardware.

For operational, backup, and data security reasons, the IRIS DMC makes five copies of each sample of waveform data. Data are stored in a time and station sorted order to optimize servicing of data requests, a second copy of each of the time and station sort orders is also archived for redundancy, and one copy of the data is stored off-site on DLT tape for safekeeping. These safeguards effectively increase the archiving requirement by a factor of five, making our mass storage system requirement 30 terabytes per year and our offsite DLT storage another 7.5 terabytes per year. There is no intention to change this basic data archiving strategy.

The current mass store system at the DMC has an installed capacity of 180 terabytes and with modular expansion can be increased to 360 terabytes, sufficient to service all USArray data in addition to existing data sources. The incorporation of higher density tape drives can increase this capacity to more than one petabyte.

Data Distribution and Archiving

The building and handling of the data products for USArray waveforms will be based on a variety of archiving and distribution capabilities that have been developed over the past 15 years to serve the needs of the research community. The primary goal is to provide users with a complete and continuous archive of quality controlled information (waveforms and associated metadata) from all USArray instal-

	Contributing Sites	Number of Channels	Sample Rate, Hz	Duty Cycle %	Data Rate KB/sec	Data Rate MB/day	Data Rate GB/yr
Transportable Array							
Broadband	400	3	40	100	375	3955	1410
Long-Period	400	3	1	100	9	99	35
Flexible Array							
Broadband	200	3	40	90	169	1780	634
Long-Period	200	3	1	90	4	44	16
Short-Period	200	3	100	90	422	4449	1586
High-Frequency	2000	1	250	2	78	824	294
Permanent Array (new GSN)							
Broadband	13	3	40	100	12	129	46
Long-Period	13	3	20	100	6	64	23
Ultra-Long Period	13	3	1	100	0	3	1
Permanent Array (new and existing NSN)							
Broadband	27	3	40	100	25	267	95
Long-Period	27	3	20	100	13	133	48
Total					1114	11748	4188

Table II-1.1: The amount of data estimated to be produced by USArray components is shown in the above table. A total of 4.2 terabytes per year will be generated by USArray. The total rate of generation of USArray data is 1.114 megabits/second with the Transportable Array generating 384 kilobits per second. Since Bigfoot will most likely be telemetered in real time to the ANF and the DMC, these are realistic rates to achieve. The values assume average compression of the data to 1 byte per sample. The Transportable Array will consist of 400 instruments recording 3 channels continuously at 40 samples/second and 3 channels continuously at 1 sample per second.

lations. In developing this complete archive, two pathways have evolved to serve the most common requests:

- Event windowed vs. Continuous. Many seismological investigations are based on analysis of all available data from specific events (earthquakes or explosions). Once the origin information (location and time) of an event is known, simple tools can be used to extract the time windows of interest for waves arriving at any seismic station. Since these data segments represent a small fraction of the total archive, they can be stored in on-line disks for rapid access. At the IRIS DMC, these on-line resources have been called FARM (Fast Access Recovery Method, for quality controlled data from the archive) and SPYDER® (for access to near-real-time data from events, before complete quality control). Since it takes time (minutes to weeks) to create event catalogs and collect data from all stations, these on-line data resources grow with time following an event. This is especially true for the FARM archive, which depends on the completion of quality control procedures.

- Immediate vs. Quality Controlled. In general, most research experiments look for the highest quality, most complete data available, In the case of the DMC, the resource of choice is the permanent archive of continuous data, or the FARM for event-windowed data. There are applications, however, especially in earthquake monitoring and education, where immediate access is more important than completeness or final quality control. To service these types of requests, the IRIS DMC, in collaboration with the USGS, has developed a variety of user tools that

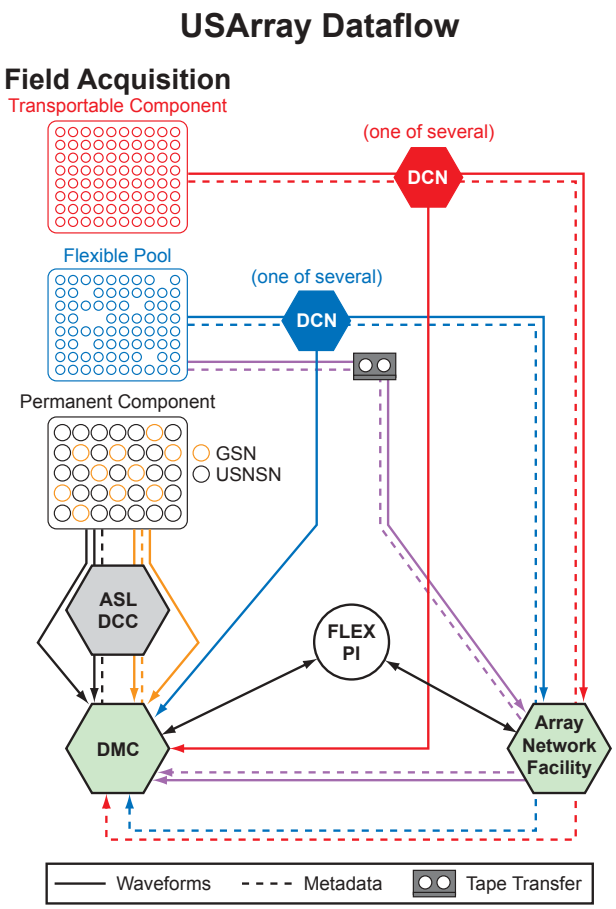

Figure II-1.9: Data will flow from the various USArray components to the Array Network Facility and to the IRIS DMC. Data Concentrator Nodes (DCNs) will forward data simultaneously to the Array Network Facility and the DMC for both the Transportable Component and the Flexible Component when real time connections are possible. Tape based transfer will flow through the Array Network Facility node and then to the DMC. Metadata generation will be the responsibility of the Array Network Facility. Backbone Network data will flow in real time to the USGS facility in Albuquerque ASL and to the DMC.

collect event-related waveforms immediately following notification of an event by the National Earthquake Information Center (NEIC). The core of this system is SPYDER®, which uses the NEIC location information to determine the appropriate time segments and gathers waveforms from stations that are available either on-line or via dial-up modem.

The USArray data will be processed and stored in a manner compatible with the way in which all other data are managed at the DMC. Waveform data entering the DMC will be handled using well-established international standards for formats and metadata (SEED and miniSEED). Procedures are in place to exchange metadata information with network operators to update needed information related to station configuration. The waveforms will be stored for four months in an on-line disk-based RAID system and the metadata will be managed in an Oracle Database Management System in a manner analogous to the way all other passive source data are archived. Data that are acquired from active source experiments will be received and stored in SEG-Y format and distributed as special volumes of "assembled data sets."

USArray Data Flow to the DMC

Data Flow From the Transportable Array

Data from the Transportable Array will be sent from the field in real time using TCP/IP communications protocols. Data from stations will flow from the stations to a Data Concentrator Node (DCN) that will be located where it can be connected to the Internet with high speed, reliable links. Often these links may be located at existing U.S. regional network data centers. From the concentrator, the data will simultaneously flow to the Array Network Facility (ANF) and the IRIS DMC. In the event that circuits from the DCN to either the DMC or the ANF fail, a dedicated frame relay link between the ANF and DMC will be used as a redundant communications path to the other center.

At the DMC, the data will be managed in a parallel system, dedicated to the management of USArray data. The data will flow into a Buffer for Uniform Data (BUD) system, similar to that used currently for reception of various data streams into the DMC. The data can then be made available through the BUD real-time data access methods and through the WILBER interface to the SPYDER® products.

IRIS is presently developing methods to automatically implement routine procedures to check data quality. Data from the BUD system will flow through these quality assurance tools and into the primary DMC archive. With about a five-week delay, data for larger events will be extracted from the archive and FARM products will be formed.

Data Flow From the Permanent Array

Data flow from the permanent stations of the Backbone Network will be similar to that for the Transportable Array but will use communications systems established for the GSN and ANSS. Since the ANSS serves an essential role in operational monitoring of national and global earthquake activity, the data collection for these stations provides for additional redundancy and includes a direct node at the NEIC.

USArray Data Flow from DMC to Users

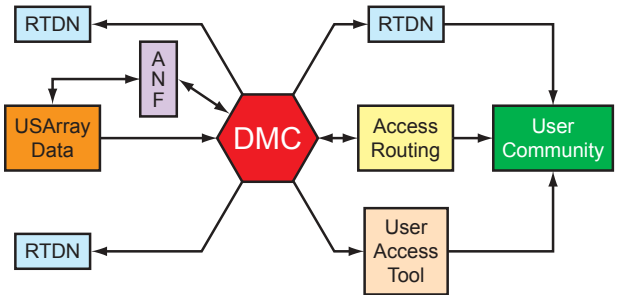

Figure II-1.10: The combined real time data rate from USArray may be more than 1.1 megabits per second (see Table II-1.1). Even data from the Transportable Array itself will be about 400 kilobits/second, one-quarter of a T1 circuit. As such it will not be likely that a single distribution point will be able to distribute all USArray data in real time. A data distribution system built upon the IRIS DHI model will be needed. The concept is to populate several Real Time Data Nodes (RTDNs) with copies of the real time data. Distribution to the community can be done from this distributed system. Real time data requestors will connect with a Access Routing server at the IRIS DMC to determine where to connect for real-time data streams. Non-real time data users will access data through traditional User Access Tools from the DMC. A hierarchical system of secondary and tertiary RTDNs could also be easily built using this system. Similar in concept to the Unidata Internet Data Distribution (IDD) System, this system would also support streaming data. See Figure II-1.11 for more detailed view of DMC archives and User Access Tools.

As the primary operator for the permanent USArray stations, the ASL will be responsible for monitoring operational status and for preliminary quality control. The data will then be forwarded to the IRIS DMC for archiving and distribution.

Data Flow From the Flexible Array

As the name implies, the Flexible Array will be deployed in a variety of sizes, station geometries, and implementation modes. Data collection will vary significantly from one experiment to another and will be largely defined by the needs of the individual experiment and the Principal Investigator. Data from the Flexible Array will sometimes be telemetered and in other experiments will be locally recorded at the station site.

In telemetry mode, the data flow will be similar to that used for the Transportable Array, and data will be sent to the ANF, where quality control and reformatting will take place, and forwarded to both the IRIS DMC and the Principal Investigators. At the DMC, these data will flow through DMC quality assurance tools and then be archived with the other DMC data.

In experiments where on-site recording is used, the resources and tools developed at the ANF and DMC for quality control and data assimilation will be used to insure uniform data quality. It will be the responsibility of the ANF working together with the PI to provide final data products for archiving at the DMC. The involvement of additional USArray resources for data collection and during the experiment, and the mode of delivery to the DMC archive, will be defined during the planning stage for each experiment.

Management of Real-time Data

In response to the increased use of real-time data collection in the GSN and PASSCAL programs, and in anticipation of the USArray portion of

EarthScope, the IRIS Data Management System began the development of the BUD system to ingest and manage large amounts of real-time data. Real-time data from seismic stations and networks can arrive in a variety of formats and via various communication protocols. The BUD converts these data into a standard format for use internally within the DMC and to provide a standardized interface to provide external users with access to real-time data. The BUD system has been functioning since 2001, and currently handles more data in real time than any individual USArray component will generate. Therefore, IRIS now has a reliable and dependable system that can receive the anticipated real-time data from USArray and we anticipate little new software development will be required for data ingestion. Not only can the IRIS DMS draw upon BUD for data reception, a series of tools have already been developed that can distribute data in real time as well. All of the systems that have been developed are scalable—as new data streams are added and as demand warrants, additional processors and RAID disk subsystems can be added to handle the increased load in a straightforward manner.

To minimize the impact of USArray data flowing into the DMC, a complete clone of the BUD system will be installed in order that the new data streams do not overtax existing DMC systems. USArray will require the installation of at least one SUN Enterprise class server and RAID system to manage USArray data. BUD applications and utilities will be installed on the new system. Since USArray will generate a very valuable scientific asset, we as-

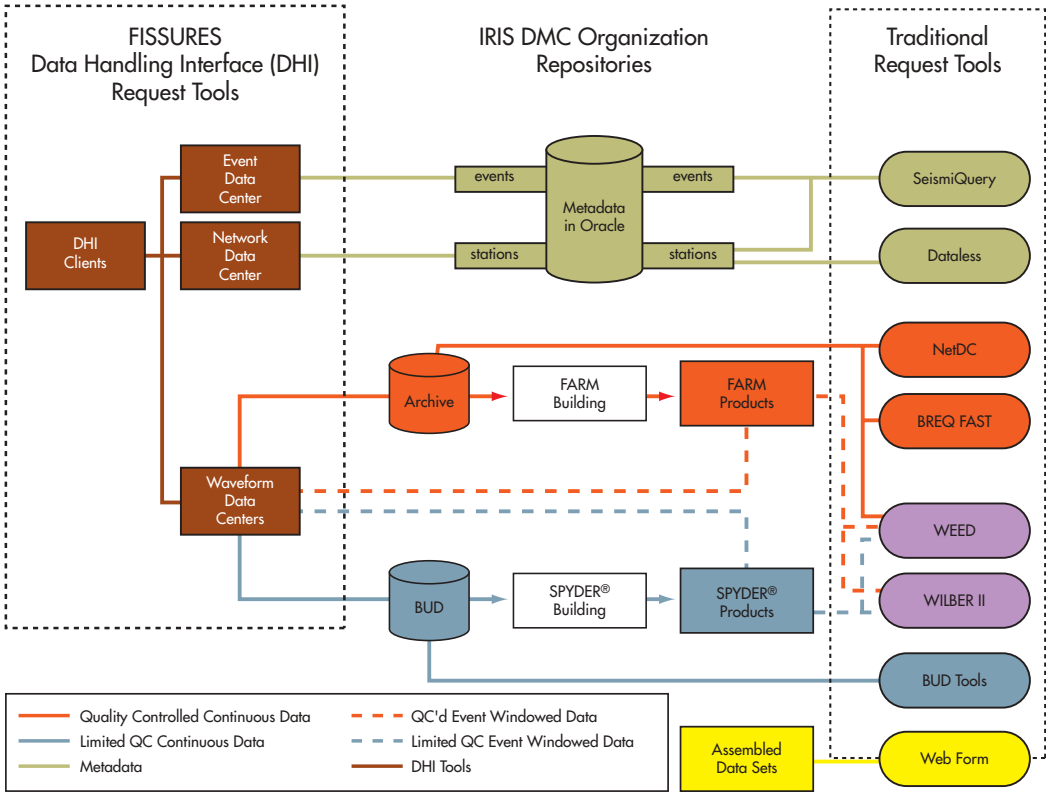

Figure II-1.11: This figure shows the four data repositories (Archive, BUD, FARM, and SPYDER®) that exist at the IRIS DMC. Continuous data are held in the Archive and the FARM; event-segmented products are in the FARM and SPYDER® systems. Data in the Archive and the FARM are quality controlled whereas BUD and SPYDER® data are real-time data with little or no quality control. USArray data will be available through a large variety of traditional data request tools supported at the IRIS DMC (as shown on the right) as well as the new FISSURES/DHI tools that support direct data access to clients from three types of Data Center Servers (Event, Network, and Waveform).

sume that users will generate an increased number of data requests for these data. To accommodate both the increased amount of data flowing into the DMC as well as an increase in the number of data requests, we anticipate increasing the throughput of the primary mass storage system by installing several more tape drives to the Powderhorn robotic system used in the DMC archive. This will allow more data to flow into and out of the primary mass storage system without degrading access to existing data sources.

Distribution of USArray Data in Real Time

To a first approximation, USArray data from the Transportable Array will generate roughly one-quarter of a T1 data communications circuit. If only a small number of users wish to receive real-time feeds of the USArray data, the Internet connection at the IRIS DMC could become a bottleneck. IRIS has anticipated this and has begun to design systems that will need to be developed for a distributed system for real-time data.

Leveraging technology developed within the IRIS DMS FISSURES project, the Data Handling Interface (DHI) has become a viable method of distributing data in real time. In order to meet the anticipated real-time data distribution requirements of USArray, IRIS will develop a distributed data system for real-time data that will be built upon the DHI. Each node of this distributed system will consist of an inexpensive workstation and a disk buffer capable of holding one week of USArray data. DHI-based software will be installed on these systems and the turnkey systems can be installed at selected EarthScope participating universities as needed. This will be an effective was to ensure all desired USArray data is available to regional networks as desired. The universities will gain the advantage of having the USArray data available immediately from the local disk buffers.

The DHI-based software will transfer complete copies of the USArray data in real time from the IRIS DMC to each of the distributed nodes. DHI servers will also be installed at each of these nodes to provide full metadata and seismological waveform data as needed. IRIS will develop an intelligent DHI routing system whereby DHI-enabled clients will access a DHI server when the client wishes to gain access to real-time data streams from USArray. The server will determine the nature of the access desired and determine which of the distributed nodes can best serve the real-time data needs of the client. The server will direct the remote client to the most appropriate distributed node to meet its real-time needs. The client will then connect with the indicated server and transfer the requested waveform data or information from the RTDN to the client machine. The development of the real-time data distribution system will require funding from the Earthscope Operations and Maintenance as it is not specifically included as an MRE effort.

For non-real-time access to data, the powerful set of standard IRIS user tools will be available to access data from the archive (see Figure II-1.11 and www.iris.edu/manuals/DATutorial.htm). We project that the centralized node of the IRIS DMC should be able to continue servicing these requests directly through the existing DMC systems. As demand warrants additional resources can be installed at the DMC to scale the capabilities to meet user demands.

Establishment of an EarthScope Data Portal

The UNAVCO and IRIS data handling systems are very mature and meet the needs of their respective communities very well. In the early stages of EarthScope activities we anticipate the two individual data management systems will continue to function well, in a manner similar to their operation today. Emphasis will be on scaling the existing sys-

tems to handle the increased data flow rather than the development of new technologies for data distribution or for integrating activities.

Of particular note within the respective systems are the UNAVCO Seamless Archive and the IRIS Networked Data Center concepts. Both of these systems are functioning and offer access to data in a distributed data center environment. UNAVCO can offer data from multiple GPS data providers in a one-stop shopping environment and the networked data centers system developed by IRIS offers access to seismological data residing in more than six different global and regional data centers.

As the initial step, we anticipate the development of an EarthScope portal through which the individual data management systems will be easily discovered and through which data are easily accessible. We will create this EarthScope data portal early in the process and will bring it on-line long before significant data from PBO or USArray begin to flow.

The development of the data portal and the support of the seamless archive and the networked data centers are the only parts of data distribution to end users that IRIS and UNAVCO propose to implement using MRE funding. All other development will rely upon Operations and Maintenance funds or funds that can be found from other sources such as Information Technology Research and/or Cyberinfrastructure.

Integration of PBO and USArray Data

The CORBA based technology driving the IRIS DHI system is complicated but extremely powerful. It falls in a class of software reserved for Enterprise applications, such as EarthScope. One simple perspective of Enterprise systems such as CORBA is that the entire system is tightly controlled and understood by everyone involved in the development of servers and clients. In general such systems are not intended for casual, infrequent developers.

CORBA requires that well-defined interfaces exist to various kinds of data and information, and that client applications can access the various kinds of information. IRIS is experienced in developing the interfaces required to access most kinds of seismological data. Development of CORBA interfaces to GPS data has not yet been undertaken. UNAVCO and IRIS will jointly pursue the extension of FISSURES interfaces to geodetic data. In so doing, tight integration of seismological and geodetic data can be accomplished such that seamless access to both types of information will be possible.

While we believe CORBA to be a viable technology, developments in this area continue to take place at a rapid pace. For this reason we will examine other technologies, in addition to CORBA, to ensure that the technology we implement makes the most sense at the time we develop it. While we believe that choice today is CORBA, our solution may be different (XML, SOAP, XSLT) when the time for implementation comes.

Once the Interface Definition Language (IDL) or equivalent schema is designed for PBO and USArray data, staff at the IRIS DMC, UNAVCO Boulder Facility, or other university locations, will develop clients that will access the servers at IRIS and UNAVCO locations. These clients will provide seamless access to the GPS and seismological observations as well as the variety of data products and derived products generated using PBO and USArray data.

EarthScope Products

The raw data from the core EarthScope PBO and USArray facilities will be large volumes of GPS and seismological waveforms and associated metadata. These primary observations are esoteric and only

meaningful to experts in the particular sub-disciplines. However, there are routine products that come directly from the observations that are understood by a much broader community. For instance, catalogs of earthquake locations or crustal velocity models and images are well understood and very useful to a broad community. Similarly, experienced users of GPS data desire access to the raw phase data but these data are not of general Earth science use. However the routine calculation of precise locations or derived motion vectors is generally understood and possesses broad application. As an initial stage in the development of higher level EarthScope products related to USArray and PBO, the IRIS and UNAVCO data systems will develop systems to manage and distribute a wide variety of derived products or packages of primary observational data (such as the IRIS FARM) for easy access to and use by the scientific community.

The exact types of products need to be defined by a broad cross section of the Earth sciences community. As these products are defined, and processing methods developed to produce them, IRIS and UNAVCO will develop data management plans for these products under the Operations and Maintenance portion of EarthScope.

Education and Outreach

All EarthScope components are committed to ensuring that the data collected by the facilities will be openly available to all interested parties—including the public and especially the educational sector. We are aware that to be useful to the K-16 educational community, it is not sufficient to simply declare the data "open" – it is essential that data be provided in formats and as products that are accessible to educators and students, and that there be appropriate teaching modules to allow the resources to be incorporated into an inquiry-based learning experience.

IRIS intends to provide educational linkages to USArray through both the existing IRIS Education and Outreach Program and through a larger EarthScope educational enterprise as it develops. Facilities and experience developed as part of the IRIS E&O Program provide a natural pathway for access to the seismological data produced as part of USArray and to resources developed to incorporate seismology into the K-16 learning environment. Broader EarthScope outreach activities (such as those proposed in the EarthScope Education and Outreach Program Plan) will place the seismological resources of USArray and IRIS in the broader context of the full EarthScope enterprise and the Earth sciences.

Seismology-related products designed for groups outside the scientific research community will continue to be developed by the IRIS Education and Outreach (E&O) program in cooperation with other EarthScope partners. Current IRIS E&O activities are targeted at audiences ranging from K-16 students to the general public, and are focused on areas where IRIS is well-positioned to make substantive contributions stemming from its strong research and data resources. Outreach to the general public includes a distinguished lecture program, and museum exhibits with real-time displays of earthquake locations and ground motion. Efforts that engage the wider education community include a range of K-16 teacher workshops, a new Educational Affiliate membership for undergraduate institutions, and widely distributed teaching modules and associated tools. Students can access earthquake locations and global seismic data from the IRIS Data Management System in near real time as well as by selecting events from the online archives. Students can also collect their own seismic data using a stand-alone, relatively inexpensive seismograph, or with research-quality broadband instruments with continuous network connections. Consortium members are currently developing new visualization tools and classroom activities using

seismic data, and this will expanded to include data from USArray. Visualization tools include large-scale efforts like the Global Earthquake Explorer, which is being developed by IRIS E&O through a subcontract to the University of South Carolina and will provide an online seismic data analysis environment tailored to general public and educational users

As part of USArray, IRIS proposes to carry out the following core level education and outreach activities:

- Integration of USArray data into the educational data streams available from the IRIS DMC. This integration is a relatively straightforward task since most of the necessary resources are already under development.

- Interaction with the broader EarthScope E&O effort as it becomes defined. We anticipate that this will take the form of collaboration with the efforts of the EarthScope facilities, a variety of local educational alliances distributed across the country, and numerous partners representing national and local organizations with similar science, education, and outreach goals.

- Development of both generic and region-specific outreach materials related to USArray activities as the array installation proceeds across the continent. The deployment process will provide unique opportunities to introduce local residents to seismology and the Earth sciences. For example, the museum and distinguished lectureship programs, which currently are focused on large museums throughout the US, could effectively be expanded to also target the smaller communities where USArray stations are located. Providing educational seismographs to schools may be able to help to play a role in permitting USArray sites by establishing connections to local communities before the arrival of USArray. A short video designed for landowners and park officials could be produced to describe the purpose and requirements of a USArray site. All of these activities will be closely coordinated with UNAVCO in the deployment of PBO.