Data Services Policy Related to the Acceptance of Large Datasets Generated from Temporary Deployments (LDTD)[1]

Policy Version 1.2

BoD Approval on May 2016

Motivation: The IRIS DMC has historically accepted most data sets offered to it at no cost to the data generator. We were able to do this due to adequate support provided by the National Science Foundation and the rate of data ingestion being moderate without significant impact on resources at the DMC. In a 2012 External Review of Data Services, the need to have a well-crafted "Data Acceptance Policy" in place that clearly articulated the requirements and responsibilities of both the data contributor and the DMC was identified. The existing Data Acceptance Policy for Passive Data Sets (DAP) was developed and approved by the IRIS DSSC but primarily addresses moderate-sized passive networks that would generate continuous data volumes of less than **10 terabytes**[2] annually.

For the purposes of this LDTD policy a large dataset is one that will generate more than 10 terabytes of data in a 12-month period. Normally these experiments will last less than one year but could last a few years.

As IRIS researchers begin to display an interest in new active/passive experiments capable of generating very large volumes of data in formats other than the traditional SEGY or SEED, it is clear that a new policy is required to insure that IRIS has the resources to manage LTDT datasets in perpetuity. LTDT experiments could result from large numbers of instruments recording data over a short period of time or from fewer instruments recording data at high aggregate sample rates or both.

Preconditions:

1. Provider must sign a ***Data Provider Agreement*** with IRIS that outlines responsibilities for each partner and details the services provided by the DMC.
2. **Coordination between the PI and Data Services:** Before submission of a proposal or significant planning of a short-term data acquisition activity that will generate significant amounts of data, the data needs of the experiment must be clearly coordinated with IRIS DS. Things the PI must identify include but may not be limited to:

    The number of stations and channels
    The data generation volumes of the individual channels
    The format in which the data will be delivered to the IRIS DMC
    The anticipated duration of the experiment

[1] This document will be reviewed annually at spring meetings of the IRIS Data Services Standing Committee
[2] In version 1.2 the threshold to trigger the LDTD policy has been raised from 5 terabytes to 10 terabytes to reflect the reduced cost of storage. Figures in the following table were also adjusted.

The dates when the PI anticipates data will begin and end being delivered to the DMC.
The method by which the data will be sent to the DMC
Electronically and if so by what protocol
By physical transfer
The total volume of the data set
The source of the funding for the data acquisition.
Any resources, in addition to storage capacity, the DMC will need to manage the data including conversion of the data from the submitted data set to one of the DMC supported formats (currently SEGY or SEED).
The openness constraints on the data (how long might the data be restricted if at all)

Coordination Between IS and DS: It is anticipated that the PASSCAL Instrument Center will be involved in many but not all of the LDTD experiments. In order to insure close coordination between the two Directorates of IRIS a web application will be developed and jointly maintained within which all of the key information about an experiment will be maintained. Access will be granted to the PI and their designated co-PIs, PASSCAL staff, and DMC staff as needed. NSF should also be allowed access in order to identify which large data generating projects have been funded.

Cost Recovery by the IRIS DMC. The cost to manage data from LDTD experiments is not included in the core funding the DMC receives through the cooperative agreement between IRIS and the NSF. For this reason, larger data acquisition activities must include the data management costs in the proposal to their funding agency, for costs that will be incurred at the DMC. This cost recovery is a **one-time cost** and is limited to the costs the DMC will incur getting the data into the DMC's systems. The DMC will assume the costs for the on-going management of these data within its existing budget for the out years.

For projects funded by the NSF the costs to be recovered include the following:

- Direct costs of the storage systems required to manage the data sets at the IRIS DMC, at the Auxiliary Data Centers (ADCs), and for the tape backup systems at the DMC and the ADC. This is currently (2016) set at $1250 per terabyte. This covers the hardware that will be used.
- Ingestion costs such as conversion costs from non-traditional formats or data ingestion at the DMC at a cost of $250 per terabyte. This covers the cost of Data Technicians to conduct the format conversion.
- Metadata creation for data to be managed in SEED format if the PI does not wish to generate the metadata themselves
- Other costs to deal with specific issues that are unique to the LDTD Data set in consideration. This might include such costs such as development of new conversion software, generation of metadata to enable conversion to either SEED or SEG-Y format or other new capabilities.

We assume that various experiments will wish to use at least one and perhaps all of the above possible services. Costs for the above services are shown below. These are intended to be one time costs.

Task	Cost	Per Unit	Cost Basis
Management of data by the DMC	$1,250	terabyte	Cost of storage hardware
Ingestion costs requiring reformatting	$250	terabyte	Cost of Data Control Technician (DCT) staff time
Creation of metadata	$2,500	100 channels	Cost of DCT time to create metadata
Support for non-standard formats other than SEED (dataless and miniseed)	TBD	Per experiment	Development cost, dependent on the specific capability development cost

Enhanced Recovery Costs

The National Science Foundation covers the majority of costs of developing and operating the IRIS DMC. The EAR Divison within the GEO Directorate is the direct source of this funding. The cost of this infrastructure is significant and as such an enhanced recovery cost to manage data for non-NSF related LDTD experiments is being factored in to acknowledge the role that the NSF funding plays in the normal operation and development that takes place within the DMC.

	Enhanced Recovery Factor
NSF and non-profit educational/research institutions	1.0
Data sets generated by other Federal or State Agencies	2.0
Data sets generated by for-profit organizations wishing to use the services of the IRIS DMC	3.0
Non-US non-profit organizations with sustained operations	1.0

NSF and US non-profit educational/research institutions

Projects that are NSF funded or funded by a non-profit educational or research institution. This includes both IRIS members and non-IRIS member universities.

Other U.S. Federal or State agencies

This category includes other federal agencies that may wish to use the services of the IRIS DMC. It could include the following federal agencies (Interior (USGS), Energy (National Labs), State, Commerce (NOAA) as well as other federal agencies requesting DS services.

For-profit companies

Examples of these organizations would be US or international oil companies, mining companies, or geotechnical companies.

Non-US non-profit organizations with sustained operations
The data contributions that other countries make to the IRIS DMC's data holdings are scientifically very significant. IRIS's ability to host their data contributions offsets costs that do not need to be borne by the network operator. Examples of such non-US organizations with sustained funding include the Geological Survey of Canada, GNS in New Zealand, Geosciences Australia, the SEIS-UK program in the United Kingdom.

Rather than asking direct contributions to IRIS and given the potentially high scientific value of data generated by these networks, but also recognizing the cost savings realized by these networks, we suggest that these networks contribute to a fund administered by IASPEI but managed by the FDSN to support training and infrastructure support in the developing world. These funds will support making turnkey network operating systems available, support training activities in various parts of the world, and supports acquisition of equipment for networks that will result in data being made available to the broader global seismological community. The requested contribution rate should be at the NSF/GEO/EAR rate.

These funds will be used to enhance capacity building in the developing world by supporting international training workshops, programs to make data available to and from the developing world, and improvements to seismic network operations software.