# LTER Data Accessibility: Barriers and Solutions

LTER Network Information System Advisory Committee, May 2012

The LTER Network has adopted a policy of open access to data and has expended enormous effort by both sites and the Network to achieve that ideal. At present, the LTER Network makes over 6000 data sets directly available through a central data portal, and many more data sets are accessible through site web pages. Many LTER data are also available through secondary portals such as the ORNL DAAC, the KNB Metacat, and other repositories. Additional LTER data are well-documented and managed by sites, but are embargoed under Network data policies because they contain information on endangered species, critical habitats, or graduate projects that are not yet completed. Overall, the LTER Network has made significant progress towards the goal of full data accessibility and continues to advance towards that goal.

The LTER Network Data Portal currently provides metadata for 6,823 datasets (LTER Data Portal Browse Page). Within these metadata documents are 11,372 network links (query of LTER Metacat 3/22/2012). Of those links, 6,527 refer to specific data resources (i.e., “entities” such as data tables, GIS data, etc.). An additional 4,845 links, at the dataset level, provide less-specific links to data resources. In addition to those found in the LTER Data Portal, individual sites may provide access to additional data resources on their own web pages.

Despite the availability of many LTER data sets, the recent 30-Year Review has directed attention to those data sets that are not yet accessible. To address concerns about these data, the LTER Network must make a concerted effort to identify barriers to making additional data accessible, develop solutions to overcome these barriers, and marshal or acquire resources to implement these solutions. As part of this effort, the LTER network needs to identify and implement metrics that quantify progress toward the goal of full data accessibility. In this white paper we provide a summary of barriers to data access, a list of possible solutions, and the resources required to implement them, and discuss several issues, including whether LTER should record the identity of data users and ways to track LTER progress with respect to achieving the network goal of maximizing the availability of LTER data. Finally we provide three recommendations for action by the LTER Executive Board and Science Council.

## Identification of Barriers

The ultimate test of data accessibility is if data from LTER sites can be easily discovered and downloaded from a single portal. Ideally, users should be able to perform a search, evaluate the metadata results, and download data directly. Today, to comprehensively locate LTER data on a topic, the data searcher must search both site data catalogs and the LTER Network Data Portal, and must be prepared to follow an often circuitous route to the data itself. This reality creates a dichotomy of actual and perceived barriers that prevent discovery, evaluation or download and result in a failed attempt to access data. In addition, the joint focus on site data catalogs and the LTER Network data portal creates uncertainty about the proportion of LTER data that are accessible online.

Data are inaccessible online if they are not described by metadata, are without a link to data, or are embargoed. Even accessible data may be perceived as inaccessible if they are unavailable through designated links, misidentified, poorly titled, badly described, or hidden behind unfriendly user interfaces or authentication barriers. If data are accessible from LTER web pages, but not from secondary web sites, they can also be perceived as inaccessible. LTER needs to address issues of both actual and perceived inaccessibility. The ultimate goal is to demonstrate high data accessibility through a single LTER Network portal.

### Actual Barriers

There are three types of "actual barriers" to data access. First, data collected, but never archived or included in a data catalog or portal, require significant effort to be made accessible because metadata must be written (often without the creators' assistance), and data must be converted to a form suitable for electronic storage. This can be a lengthy and expensive process.

Second, datasets may be described by metadata but lack any link to the underlying data. Whether by act of omission or commission, this is a case where the metadata are searchable and the data are known to exist but cannot be accessed. The metadata may provide contact information to the creator or manager, facilitating eventual access to the data.

Finally, data may be embargoed for legitimate reasons such as stipulations of funding agencies or the desire to provide graduate students time to publish their data before public release. In many cases, metadata are available for these data but do not include a link to the data. Some fraction of embargoed data will eventually be made openly available, but other data will not. The existence of such permanently restricted data must be strongly justified and a notice to that effect included in the metadata.

### Perceived Barriers

Perceived barriers to data access are diverse. In many cases, users locate metadata but do not easily locate an existing data link, preventing data access. In other cases, users locate metadata that are poorly identified, vaguely titled or that fail to adequately describe the available data. In the latter case, users may elect not to download the data because they are unable to tell if the data contain the information sought. Users may mistakenly expect to find data collected at an LTER site but not collected by LTER. Finally, high-quality metadata may contain a link that does not lead directly to the data, instead leading the user to a poorly-developed user interface or an authentication barrier that effectively discourages access to the data.

The lack of uniformity among site data catalogs, which leads to a similar lack of uniformity among data within the LTER Data Portal, is the major source of misperception about data accessibility. Given that heterogeneity among sites is embedded in the cultures and information management approaches at sites, it is doubtful that issues of perceived inaccessibility can be addressed at the root cause. Accessibility through a central data portal, while not an insignificant challenge, is still more tractable than requiring sites to refit to a common data management model. To achieve the goal of data accessibility through a central data portal, the LTER Network will have to address several issues that lead to the perception that data are inaccessible. A summary of barriers, their severity, and possible solutions are listed in Table 1. Solutions that NISAC considers to be of the highest priority are shown in boldface.

## Table 1: Issues Accessing LTER Data Online and Possible Solutions

| Barrier | Severity | Possible Solutions |
|---|---|---|
| A. Locations of download links are inconsistently placed in metadata or inconsistently displayed in the LTER Data Portal | Mild – users may miss seeing some data links | 1) **Code all links to data consistently in the "Entity" portion of EML**<br>2) Alter stylesheets to report links from different locations in EML in a consistent place within the catalog listing |
| B. Restricted Data (e.g., data from current or recent graduate students) is not accessible | Moderate – access to data is delayed | 1) Need to provide notification<br>2) Develop best practices and training of students (it is often difficult to get good metadata from graduate students) |
| C. Not all data available on LTER Site web sites is also available via LTER Data Portal. | Moderate – impedes discovery of cross site data | 1) **Mandate that data downloadable at sites also be included in the LTER Data Portal**<br>2) Establish quantifiable measures of progress |
| D. Some data links point to directories, or query interfaces - not specific files | Moderate – data can't be automated | 1) Combine multiple data files or tables into single (large) data structures (e.g., zip)<br>2) Develop web services that can enable query and/or selection |
| E. Some short-term data are in non-tabular, non-standard data structures (e.g., graduate student spreadsheets) that may be difficult to document or use and may become unusable due to format obsolescence | Moderate to severe – data accessibility and use cannot be automated, and data may ultimately be lost | 1) **Provide training for graduate students and investigators regarding desirable formats for archival data**<br>2) Hire additional Information Management personnel to work on extracting and reformatting data |
| F. Some LTER Site web sites have (temporary) technical problems that prevent downloading of data | Severe – data links cannot be followed | 1) Periodic checks (automated) to make sure links in EML are kept up-to-date<br>2) Ingestion of data into PASTA to provide an alternative download source |
| G. An unknown quantity of LTER data is not in either site or network data catalogs | Severe – "dark data" cannot be discovered or used | 1) **Charge site PI's and IM's with rigorously enforcing LTER Data Policy**<br>2) Review papers and proposals coming from sites to identify "missing" data |

## Possible Solutions and Required Resources

In general terms, these are things we all can do to improve LTER data accessibility:

**What scientists can do:**

- provide timely, descriptive metadata to information managers
- confirm that data sets are represented properly at the site and the network
- actively engage in the discussion about short-term solutions and NIS design discussions

**What the sites can do:**

- improve metadata by providing better titles, expanded abstracts, standard keywords, details about tables and consistent placement of network links in conformance with recommendations of the LNO and LTER Network Chair (see below)
- bring metadata into conformity with the EML Best Practices version 2
- bring site-based data accessibility to a consistent level with “admired” sites

**What the LTER Network Office can do:**

- perform quarterly checks of Data Portal, looking for accessibility issues and non-conformities
- incorporate accessibility and usability improvements into NIS interface

**What we can do jointly:**

- develop and prioritize criteria for Congruency Checker (Scientists, IMs, LNO)
- advance development of Congruency Checker (IMs, LNO)
- review Network data policies to make sure they foster data access

In specific, five improvements to LTER metadata have been suggested by the LTER Network Chair and the Network Office to improve data accessibility and utility (Table 2).

## Table 2: Increasing Searchability and Utility of LTER Metadata

*The first three features are used for full-text searches. Titles and abstracts should also be designed for human readability. These improvements can be made immediately by sites and are not predicated on policy changes that might be pursuant to the “Recommendations” below.*

1) **Titles** - The dataset title should be descriptive, mention the data collected, geographic context and research site (what, where), and possibly, the time frame (when).

2) **Abstract** – Include an abstract rich with descriptive text, analogous to a paper’s abstract. Taxonomic information may be appropriate. This is a good place to indicate whether the dataset is ongoing or complete. Some general terms regarding methods, instrumentation or measurements should also be included.

3) **Keywords** - Since keywords are searched in LTER queries, include meaningful set of keywords identifying the LTER site and research context, a set for keywords from the LTER controlled vocabulary (http://vocab.lternet.edu), and a set for the LTER core research areas. This is also a useful place to add additional terms that do not fit into data package titles or abstracts.

4) **Data Table Description** - data table descriptions for tabular data should be complete, including attributes, and physical format with a data distribution URL.

5) **Data distribution URL** - This URL to the data is located with the data table description above. The URL should deliver a data stream and not point to another application or web page. If sites have data use forms, they should be by-passable by the network portal. Web views reflecting LTER data availability are optimized for a download URL at this location.

Resources required to implement both the general and specific actions vary widely. Some actions, such as development and adherence to best practices, are relatively inexpensive. For example, a renewed commitment by LTER lead PI's to assure that site data are being shared in conformance with the LTER Data Access Policy is relatively inexpensive. Other solutions vary widely across sites being essentially "free" for sites already in conformance, of minimal expense to others where metadata is managed primarily using database software and expensive, and difficult for sites where each metadata document must be individually vetted. Development of software at the network level can be expensive (in dollars), while providing cost and work savings for sites.

Recent NSF supplements have included optional funds for improving data and metadata, and ARRA funds are being used in the development of the PASTA framework. Although implementation of PASTA will require substantial effort by sites to prepare data for conformity, it will address and solve most of the "perceived" problems listed above by delivering both data and metadata within a single system.

## Issues Requiring Extended Discussion

NISAC has identified several issues that require more extended discussion. These are:

### Data User Identification or Tracking

A persistent and contentious issue is whether or not users of data should be identified. Typically, such identification requires the data user to either login or fill out a web form for at least the first download in a session. The basic question is whether the advantages of requiring (or allowing individual sites to require) identification outweigh the advantages of providing immediate data access to a user.

Existing uses of identification include: 1) informing data contributors about who is using their data, 2) distinguishing between different types of data users (e.g. LTER vs. non-LTER, research vs. education) for reports to NSF and others, 3) contacting data users regarding possible collaborations, 4) contacting data users regarding corrections to datasets they downloaded, 5) "customer surveys" of data users regarding any problems they had using either the data systems or the data itself, and 6) acknowledgement by data users of intellectual rights of data providers.

Currently there is a heterogeneous mix of identification procedures used within the LTER Network. Some sites require no identification for downloading of data. Nine sites use the Data Access Server, implemented by the LTER Network Office, which uses "cookies" so that a login is only required for the first download in any session to allow unrestricted data downloads from any of the participating sites. Finally, there are a number of site-specific identification forms or login systems that allow access to data from a single site.

Below is a list of advantages accruing to different levels of identification:

- **No identification** – data are available simply by following a link. No information regarding who downloaded the data other than that provided by the web site log (i.e., date, time, IP address, browser type) is recorded.
    - Advantages
        - Ultimate in simplicity
        - No barriers for users to access data (An informal study at MCR LTER found that 40% of users who clicked to download data failed to complete the download when confronted with a form. This percentage includes access by web robots.)
        - Allows search engines to download and index data values
        - No privacy issues regarding network "cookies" or other identifiers
- **Identified** (broken down by type of identification or authorization)
    - General Advantages (applies to all)
        - Allows identification of the data user
        - Data users can be notified by data providers about corrections or updates to data they previously downloaded
        - Data providers can be informed about who is using their data
        - Datasets used by classes of users (research, education, policy, outreach) can be identified for reports
        - Can separate "internal" from "external" users for reporting purposes
        - Permissions can be granted to automated programs/workflows to download data without form intervention
        - Acknowledgement of Intellectual Rights of data providers and acceptance of Data Use policy attributable to specific data user
    - Advantages for "One-time Registration, One-time-per-session Login (e.g., LTER Data Access Server)
        - Lowers barrier – Login is required only for the first dataset retrieved within a session. After that, all data are downloaded directly upon clicking a link (similar to "None" above), but knowledge of who downloaded the data is retained
        - Allows creation of different classes of users, each with different search and download privileges
    - Advantages for "Register for each download session"
        - Facilitates collection of information regarding the purpose of specific data downloads on a dataset-by-dataset basis
        - No need to remember passwords
- **Voluntary Identification** – Users can decide (separate from the download process) whether they wish to be identified by filling out a form or logging in prior to downloading data.

- Advantages
  - Allows follow-up and notification for selected users

NISAC members were polled and the majority of those who responded favored either open access to LTER data or the use of a voluntary identification system. NISAC acknowledges the importance of providing credit to data providers and facilitating engagement between data users and data providers. Servicing those needs deserves additional discussion, and there may be effective solutions that do not require user identification as part of data discovery and access. Note that the current LTER Data Portal and most site systems currently have few provisions for supporting voluntary identification, so implementing voluntary identification would require additional resources.

### Externally-funded Data

Some LTER sites include in their data system metadata and data from associated projects, but those data are not necessarily shown on the LTER Data Portal. Additionally, the data may not be accessible. This creates confusion among users and hampers searching and/or requires permission to access.

Possible solutions are to: 1) Develop best practices or revise the LTER Data Policy regarding the display of externally-funded metadata containing no data links, 2) require that new data collection efforts agree, as a condition of using LTER resources, to abide by the LTER Data Policy, or 3) continue existing practices but provide clear notification of why data are not available.

### Metrics of Success

One challenge for the LTER Network is developing metrics that will allow us to assess our success in sharing data. Some seemingly simple metrics, such as number of datasets, are actually extremely complicated because the granularity of datasets varies widely across LTER sites and the ecological community in general. For example, many sites lump all the meteorological data collected at the site into a single dataset spanning multiple years and stations. Other sites provide the data for each year or station as a separate dataset. Still others use a separate dataset for each combination of station and year. All of these approaches have some value for specific users, depending on the scope of their study, but the diversity of approaches makes simple counting of datasets, especially in comparing sites, a misleading metric. Volume of data also is unreliable. Some relatively simple instruments can yield massive data files, whereas some small datasets, such as chemical assays, may be very difficult and expensive to collect, but require little volume of storage.

Below is a list of metrics that might be applied:

1. Have each LTER site assemble a comprehensive data inventory (based on all previous proposals and publications) and track the percentage of the datasets in the comprehensive inventory that are available through the LTER Data Portal as a measure of success in sharing data. This process will help identify the “dark data” (data with no metadata). This methodology has been used with some success at the CWT LTER site.
2. Track the number publications using LTER datasets. Currently this is extremely difficult to reliably track. However, if LTER initiates use of Digital Object Identifiers (DOIs) for

data, such tracking can be made much easier because it will allow automatic scanning of journals for relevant data citations. Current plans call for DOIs to be associated with all data in the PASTA system when it is fully implemented. If it were a LTER network priority, it would be possible to implement DOIs sooner.

3. Until data DOIs are available and start to be widely cited in journals, the number of times a dataset is downloaded is a reasonable metric. However, even it can be complicated for datasets that do not require user identification for download, because downloads may include queries by search engines, such as Google, that index, but do not use data.
4. It will also be helpful in assessing the contribution of LTER within the context of the broader ecological community by identifying "peer" databases. Who are the other organizations that are providing data to the ecological user community, and how much data are they providing relative to LTER? NISAC came up with a preliminary list.
   Peer Databases:
   - Oak Ridge National Laboratory Distributed Active Archive Center/Mercury
   - Non-LTER data in the Knowledge Network for Biodiversity (KNB)
   - Encyclopedia of Life
   - Taiwan Ecological Research Network
   - National Biological Information Infrastructure Clearinghouse (NBII) {now defunct}

## Recommendations

1. NISAC recommends that the Executive Board revisit the LTER Network Data Access Policy with the specific purpose of revising the section on user acknowledgement of data use restrictions, identification and tracking. We recommend that LTER should cease all efforts to require identification of data users as part of the data download process. We additionally recommend that LTER pursue changes to data portals that will support voluntary identification of data users in order to foster collaborations and keep users informed regarding the availability of updates on the datasets of interest. The current LTER Network Data Access Policy (http://www.lternet.edu/data/netpolicy.html) includes a list of "Data Access Requirements" and would need to be modified should the Executive Board or Science Council decide to follow this recommendation.
2. NISAC recommends that the Executive Board encourage each site to address any of the specific issues related to accessibility at their site listed in Tables 1 & 2; especially the improvement of metadata to include meaningful titles, abstracts and keywords that aid searching, and descriptions of tabular data with links to the data itself to aid use; rigorously enforce the LTER Data Policy; and assure that data downloadable at sites also be available via the LTER Data Portal.
3. NISAC recommends that the Executive Board charge sites with discovering and making available any significant datasets that are not currently in either site or LTER data catalogs, in conformance with the LTER Data Policy.