Essential features of LTER EML data packages to improve discoverability and access

All five features assume the site is able to produce well-formed EML that is compliant with the XML schema, the EML schema, and the EML ID and reference parser. More specific information can be found in EML Best Practices, and the documentation for the data package Quality Engine (also known as the EML Compliance Checker, or ECC). All items are essential for uniform presentation across the network, and Items 4 and 5 will be required for ingestion into PASTA. Items1 - 3 are used for full-text searches. Titles and abstracts should also be designed for human readability. If sites don't already have data table information in EML or some other structured form, items 4 and 5 are nontrivial, and in a different class from the 3 above. The majority of LTER data are tabular, and tables are assumed here. However, these features can be applied to other data types such as spatial data (e.g., GIS).

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).
Planned Check (ECC): The quality of a title is subjective. A simple check of length is planned, and submitters will be alerted if the data package title is less than 5 words. Current Check: conducted with Metacat query (plus processing) until the ECC is completed.

2. **Abstract** - An abstract should be informative, 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 also should be included.
Planned Check (ECC): as with titles, abstract quality is subjective, and so the abstract length is a reasonable first-level check. It is planned that submitters will be alerted if the abstract is less than 50 words. Current Check: conducted with a Metacat query (plus processing).

3. **Keywords** - meaningful sets of keywords should be included. Keywords are searched in LTER queries, so this is a useful place to add additional terms that do not fit into data package titles or abstracts. Recommendations include a set of keywords identifying the LTER site, the LTER controlled vocabulary and the LTER core research areas,
Planned Checks (ECC): alert the submitter if a) no keywords are present at all, and b) if no keyword from is the LTER controlled vocabulary is found. Current Check: The presence of a keyword node can be checked with a Metacat query.

4. **Data Table Description** – Data table descriptions should be complete, including attributes (columns), and physical format.
Implemented Checks (ECC): a) the presence of attribute descriptions, b) first-level physical descriptions (e.g., delimiters, number of column).

5. **Data URL** - The URL to the data should be located at the entity-level, not at the dataset-level. URL should deliver data and not point to another application or use page. Web views are optimized for a URL at this location.
Implemented Checks (ECC): For human use, the URL must be 'live', i.e., return intended content. For machine reading, a data URL at this location is required and must deliver data, not a form.
Note: the use of an LTER-NIS DAS proxy URL that logs downloads is machine-readable by LTER.