Potential data package quality metrics and report to IMC
IMC Metrics Working Group, May 2012

Definitions:
Metrics – measurements. These contain no value judgment.
Standards – translation of a metric into a value or relative worth

Background
The issue of improved data availability through the network catalog is prominent in a request from Scott Collins to the LPIs and in the 2012 call for supplements. Both Scott's request to the LPIs and the supplement call asked sites to give specific attention to "five essential EML features". Scott's message also stated that the Network would start collecting metrics and would report these to NSF. In response, the Metrics working group thought that the 5 EML features (as listed by Scott) could represent potential baseline metrics. However, the IMC also knows that they are not sufficient for PASTA insertion.

During the IMC's April VTC, participants asked that the EML mentors start to collect information about sites' network EML submissions. If we start on that task by tallying certain features of EML docs, then we are also collecting some basic metrics about our EML for our own use. It is not yet clear how these metrics should be managed and/or distributed – even within the IMC and network. In September, the IMC planned that quality reports from the EML congruence checker would remain internal until PASTA was in production. However, given recent interest, we should expect some level of distribution beyond the IMC to be necessary before PASTA is in production. **It is important that the IMC continues to lead development of data package metrics, including the process and timeline for distribution.**

The Metrics Working Group informed Scott that they were undertaking this project, and he asked for a summary report for the Executive Board meeting at the Science Council meeting in mid-May. Each site's IM rep will receive the report as well. The remainder of this document is a summary of how these five features were tallied for that draft, as an example of how these might comprise the first data package features that we measure.

Methods
The 5 features are below. Features 1-3 represent part of the natural language description of a dataset. Their use is somewhat subjective, but there are simple aspects of them that we can tally. Features 4 and 5 are already required by PASTA.

1. dataset title >= 5 words: This is subjective, and titles cannot be evaluated by machine. But titles are prominent and required by EML, and one feature of titles (strings) that can be counted is their length. At the Quality Engine workshop a length of 5 words was chosen, so that limit was applied here.

2. abstract: Presence/absence only

3. keyword: Presence/absence only

4. attributeList: Presence/absence. A dataTable requires an attribute list, so presence of attributeList implies that a table is described.

5. entity-level URL: presence/absence of at least one entity-level URL

Each of the five features for every EML document was scored (0 or 1), The results were normalized by dividing each tally by the number of EML documents from the site, and reporting a score for each feature as a percent. For example, a score of 75 for 'abstract' indicates the 75% of the site's EML documents include an abstract element.

Datasets were queried with Metacat queries because the ECC is still in development. To preserve site-anonymity, all acronyms were removed from the results.

Results

Average and median scores for the entire network were reported; see Report, Table 1 (reproduced here).

Table 1. Aggregated normalized scores for Network for 5 EML metadata features. 28 sites were samples (26 extant sites + NIN, LNO). Queries conducted between March and May 2012. The number of data packages queried varies by date, and ranges between 6691 and 6841. Numbers are in percent.

Feature	Median	Mean	Range
1. Title	90	79	0 - 100
2. Abstract	99	84	0 - 100
3. Keyword	100	93	3 - 100
4. Attributes	97	79	0 - 100
5. URL	72	54	0 - 100
Overall	**81**	**78**	**30 - 100**

One of the IMC's requests to the Metrics WG was to attempt to characterize sites IM needs. The results were also grouped into three categories, for the aggregation in the Report Table 2, (reproduced below)

Excellent: Overall score => 99%. Essentially all site-EML contains titles of reasonably adequate length, an abstract, at least one keyword, an attributeList and at least one URL at the entity level.

Needs help: Any one score < 50%. For more than half of EML documents, at least one of these features was not present, or in the case of titles, less than five words in length.

Good: all other scores.

Table 2. Aggregated normalized scores by group for 5 EML metadata features. 28 sites were samples (26 extant sites + NIN, LNO). Queries conducted between March and May 2012. The number of data packages queried varies by date, and ranges between 6691 and 6841. Numbers are in percent.

	Median	Mean	Range
Needs Help: 15 sites			
1. Title	75	68	0 - 100
2. Abstract	99	73	0 - 100
3. Keyword	100	88	3 - 100
4. Attributes	98	71	0 - 100
5. URL	1	25	0 - 100
Overall	**71**	**65**	**30 - 81**
Good: 10 sites			
1. Title	93	90	64 - 100
2. Abstract	97	94	64 - 100
3. Keyword	100	98	90 - 100
4. Attributes	83	85	72 - 100
5. URL	79	83	71 - 100
Overall	**93**	**90**	**83 -97**
Excellent: 3 sites			
1. Title	100	100	99 - 100
2. Abstract	100	99	97 - 100
3. Keyword	100	100	100 -100
4. Attributes	99	99	99 -100
5. URL	100	99	96 - 100
Overall	**100**	**99**	**99 - 100**

Discussion

'Entry into PASTA' and 'metrics' are two different things. This is *not* a report from the EML congruence checker, which is still in development; the data for these results were all obtained from Metacat queries. This report is preliminary to the Quality Engine checks (PASTA), and some results can reflect a site's readiness for PASTA (i.e., attributeList and URL construction). Eventually, we can use the quality engine to generate metrics.

These counts represent quality, not quantity. Sheer numbers of datasets, entities or URLs are not meaningful. However, NSF suggests that some measure of quantity will

be requested, although we do not yet have a way to quantify that yet. Also, these measurements also do not reward "rich" EML. They describe very simple entry-level features.

It is important to be aware of the limitations of this summary:

1. These are crude metrics, and some false positives and false negatives are unavoidable. These measurements do not detect
    a. Broken URLs
    b. Empty elements (completeness)
    c. Type II data (in which data are described, but no URL is included)
    d. Congruence between metadata and data
    e. Appropriate use of otherEntity
2. Data included with metadata is not counted (currently a small percentage of the total)
3. Spatial data are not well represented
4. These measurements say nothing about data package maintenance patterns or a site's local system.

The metrics group has also identified several additional features of EML datasets which contribute to the list of essential features. Following the pattern above, these additional features would help us to further assess our performance:

6. EML version = 2.1: PASTA will only be able to ingest EML 2.1 (or better), due to limitations of the 2.0 schema.

7. methods: Presence/absence only

8. temporalCoverage: Presence/absence only

9. geographicCoverage: Presence/absence only

10. taxonomicCoverage: Presence/absence only (tentative)

Currently, features 7-9 allow the user to more fully evaluate the usability of a dataset for a particular purpose. In the future, coverage elements can be machine-parsed during searches.

Conclusion

The IMC Metrics Working Group has provided a description of these essential EML features and summary metrics for features 1-5 to the LTER Executive Board for review. We recommend that the IMC further discuss the use of these nine metrics in the upcoming weeks, in preparation for more thorough review at the IMC meeting in September, 2012.