Baseline Metrics for LTER Datasets

- IMC Quality Metrics Working Group
  - Established, 2009
  - Advises development of network-level tools for site self-assessment
- Activities to promote dataset quality
  - Inter-site mentors
  - Shared internal reports (2011)
  - Featured at regular IMC meetings

5 Essential EML Features

Natural language description, searching

1. Title – 5 or more words in length
2. Abstract – presence/absence
3. Keywords – presence/absence

Data entity availability

4. Data table description – presence/absence
5. Data URL – one or more

Methods

Queries to Network Metacat catalog

- PASTA-driven tools not yet complete

Scoring

- True/false for each feature
- Normalized by site's total number of data packages
- Acronyms removed from summaries

Aggregated Normalized Scores - Network
5 EML Metadata Features

	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**

Number of sites: 28 = 26 extant sites + NIN + LNO
Queries were conducted between March and May 2012
Number of data packages queried varies by date. Range: 6691 - 6841

Aggregated Normalized Scores - Network
5 EML Metadata Features

	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**

Number of sites: 28 = 26 extant sites + NIN + LNO
Queries were conducted between March and May 2012
Number of data packages queried varies by date. Range: 6691 - 6841

Aggregated Normalized Scores - Network
5 EML Metadata Features

	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**

Number of sites: 28 = 26 extant sites + NIN + LNO
Queries were conducted between March and May 2012
Number of data packages queried varies by date. Range: 6691 - 6841

Characterize Individual Site Needs

- Excellent: site overall score => 99%

  Essentially all EML has reasonably adequate title, an abstract, keywords, data description and URL

- Needs help: site has one score < 50%

  At least one of title, abstract, keyword, description or URL is missing in at least half the site's data packages

- Good: all other scores
  - Generally, 50 - 98%

Aggregated Normalized Scores by Group
5 EML Metadata Features

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

Group Definitions:

Needs help - any score < 50%
Good - all other scores
Excellent - overall score => 99%

IMC Metrics Working Group 2012

Aggregated Normalized Scores by Group
5 EML Metadata Features

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

Group Definitions:

Needs help - any score < 50%
Good - all other scores
Excellent - overall score => 99%

IMC Metrics Working Group 2012

Aggregated Normalized Scores by Group
5 EML Metadata Features

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

Group Definitions:

Needs help - any score < 50%
Good - all other scores
Excellent - overall score => 99%

IMC Metrics Working Group 2012

Caveats

- These are crude metrics; false positives and negatives are unavoidable. We cannot easily detect:
  - Broken URLs
  - Empty elements (completeness)
  - Type II data (where a URL is not appropriate)
  - Data-metadata congruence

More Caveats

- Spatial data are not well represented
- Cannot measure data package maintenance patterns, or a features of a site's local system

Simple metrics provide

- A baseline – one step above basic metadata submission
- Preliminary checks for PASTA quality engine readiness

Future directions

- Additional checks will be necessary, and could be added soon. The WG has preliminary data from all sites for these features:
  - Methods – *required for full evaluation of data*
  - Temporal Coverage – *valuable for discovery*
  - Geographic Coverage – *valuable for discovery*
  - EML version – *workflows require 2.1*

- 9 checks more clearly distinguish sites which are “PASTA-ready” from those that are “pre-PASTA”

EML Compliance Checker

- PASTA Quality Engine builds reports on data package quality
  - *Sites can pre-evaluate their data before submission*
  - *Users receive report at download*
- Metrics working group involvement
  - *Developed the requirements (ongoing)*
  - *Designed the checks (2012 workshop)*

Available with PASTA release - version 1 - late 2012

Metrics Working Group

Site Information Managers:

- Dan Bahauddin (CDR)
- Sven Bohm (KBS)
- Emery Boose (HFR)
- Jason Downing (BNZ)
- Corinna Gries (NTL)
- M. Gastil-Buhl (MCR)
- Margaret O'Brien (SBC), chair

LNO Staff:

- James Brunt
- Duane Costa
- Mark Servilla

Ad hoc: developers and members of the EML community

IMC Metrics Working Group 2012