EDI/LTER EML Congruence Checker

ECC Working Group
2019

Introduction

System’s role in curation of data and metadata

Supporting tools and infrastructure

Motivate and measure the evolution of metadata and dialects

Background

Mid-2000s:	
1000s of datasets from the LTER available secondary use increasing Narrative recommendations	Automated processing attempted, but 1. Metadata was incomplete 2. Data-metadata often incongruent

Needed a mechanism to provide feedback: data-metadata congruence and potential usability

System's Role in Metadata Curation

Working group defined basic requirements:

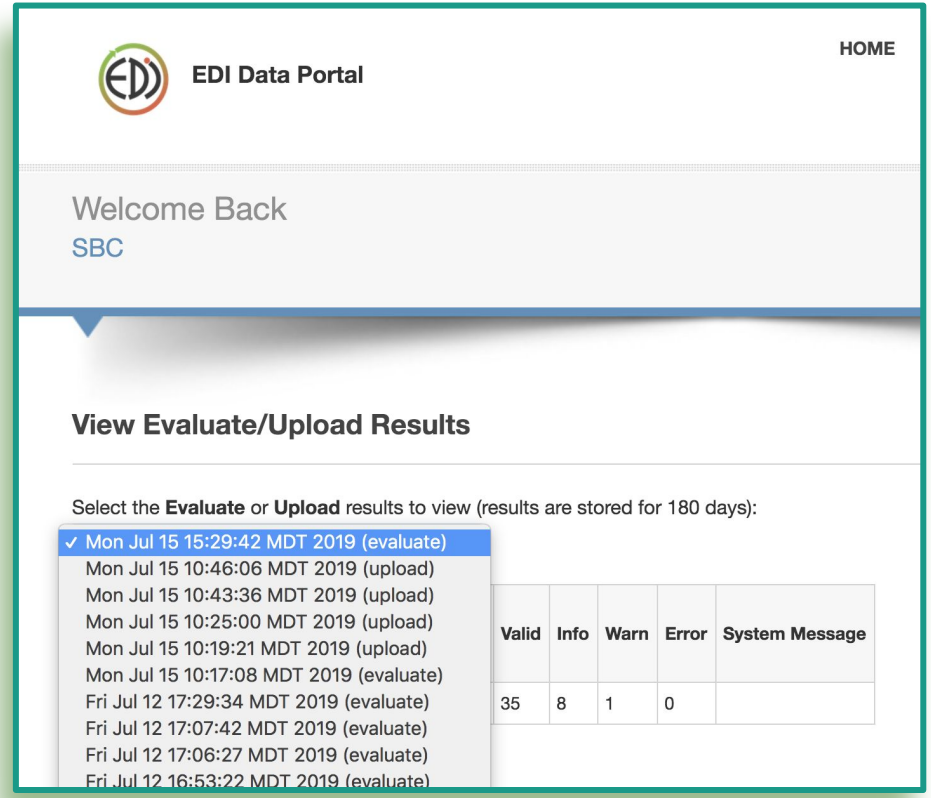
1. accommodate the addition of new checks and staged implementation
2. Configuration should be customizable for different communities
3. Checks that return "error" (prevent insertion) first
4. HTML interface for easy viewing

System's Role in Metadata Curation

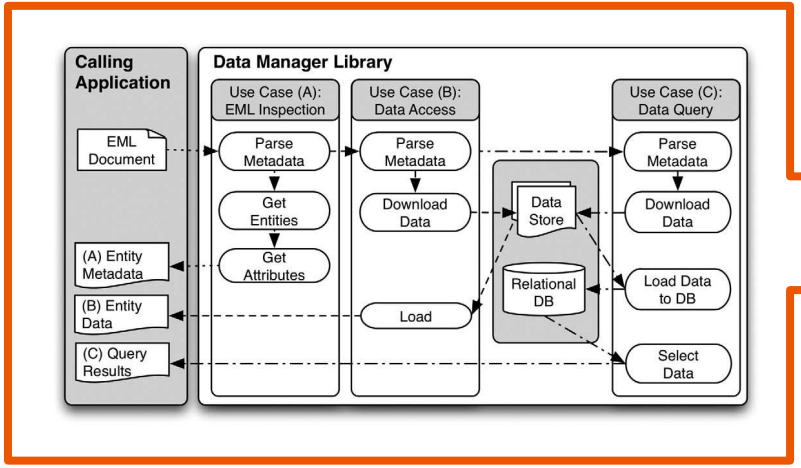
Two modes:

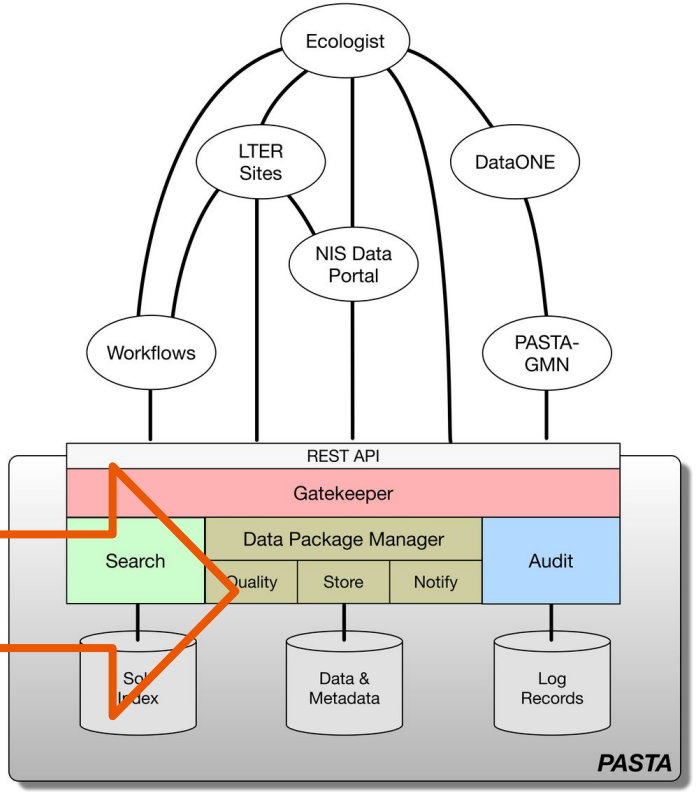
Evaluate - check as much of the package as possible

Harvest - stop on first error

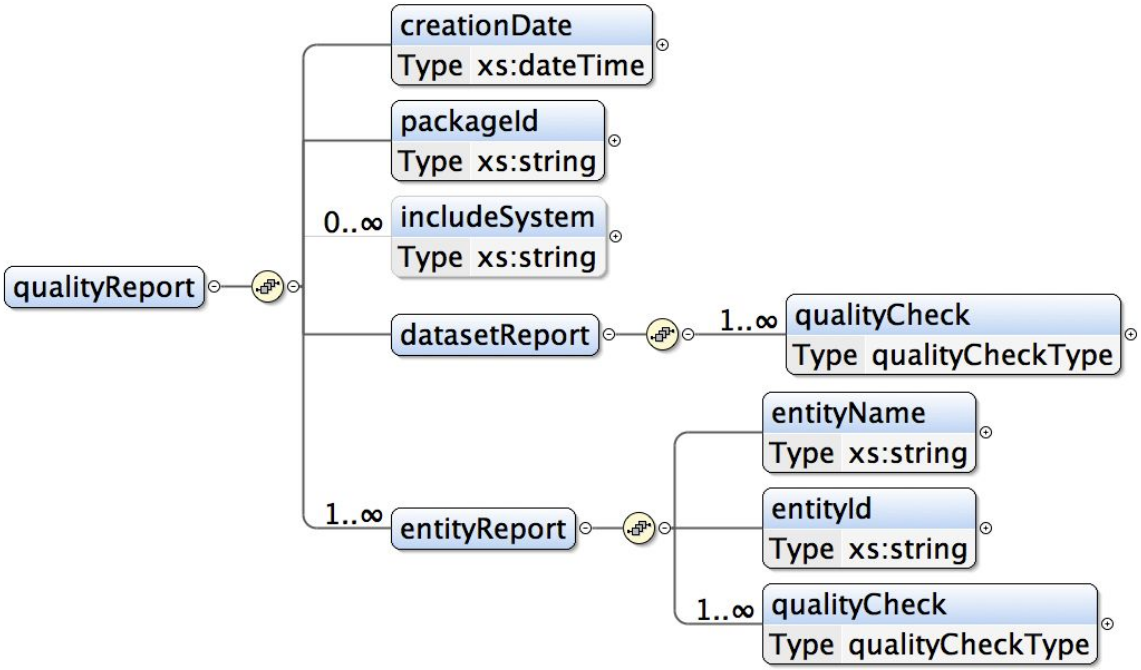


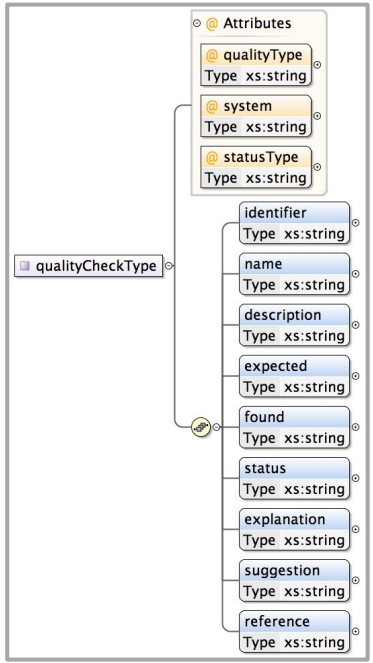
Supporting Tools and Infrastructure





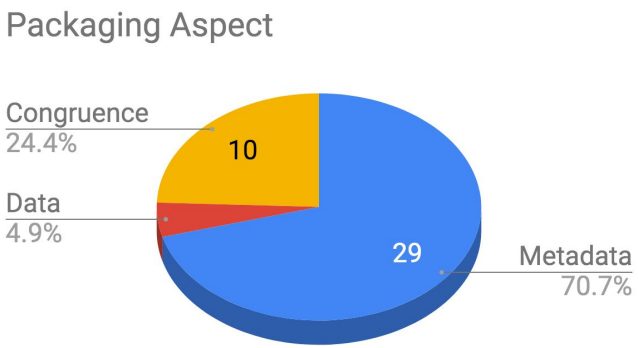
Supporting Tools and Infrastructure

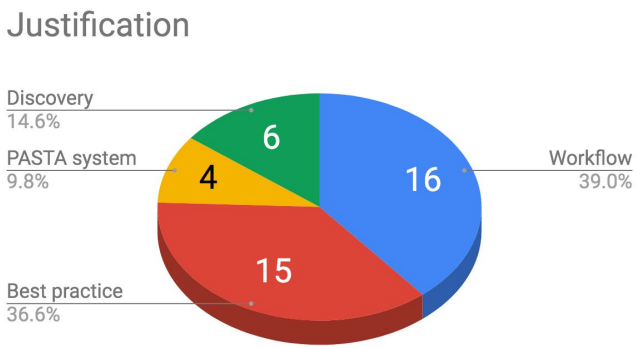




Supporting Tools and Infrastructure

41 Checks categorized according to multiple typologies

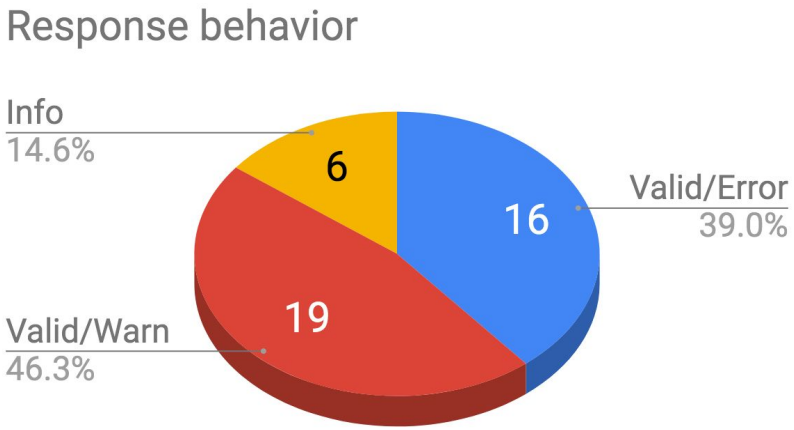




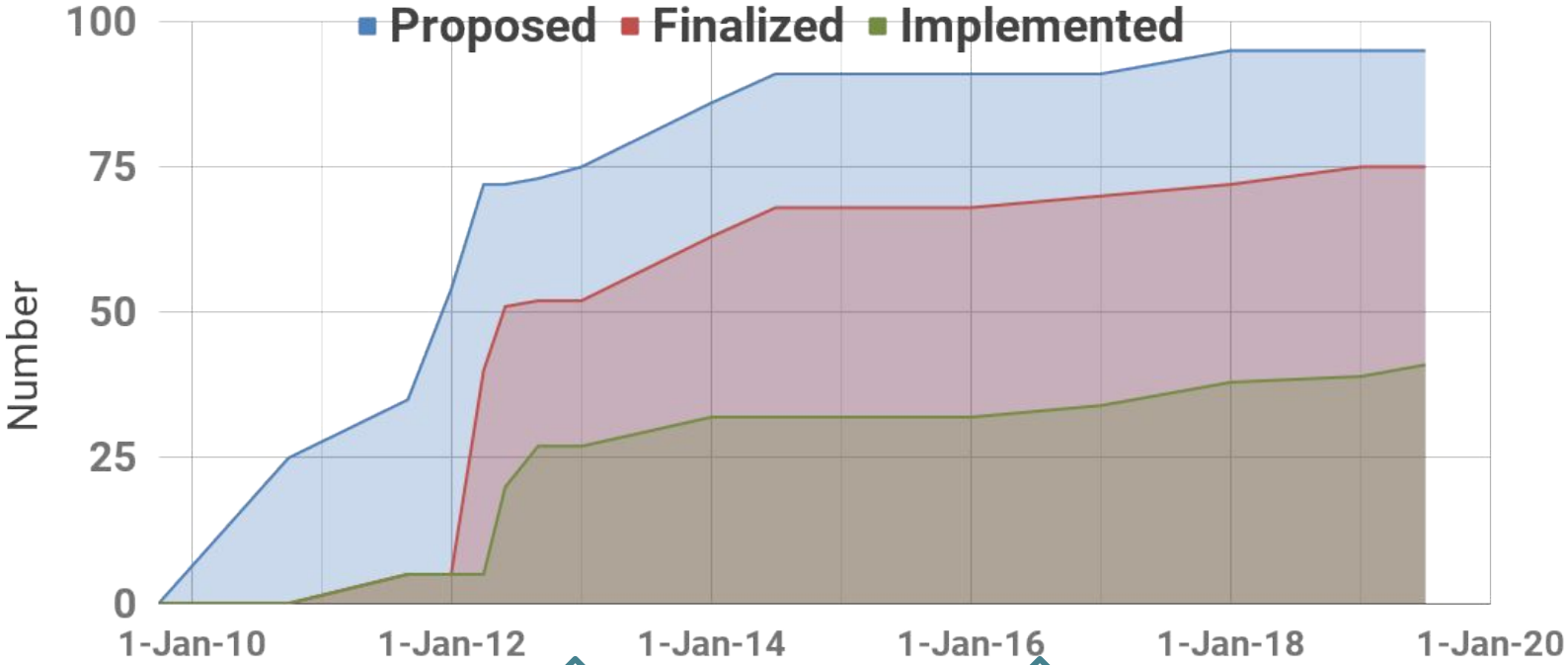
Supporting Tools and Infrastructure

Datasets are rejected only if package is UNUSABLE:

- *Schema valid*
- *Valid package identifier*
- *Resolvable data URL*
- *Checksum &/or file size congruent*
- *Unique entity names*
- *No jagged tables*



Check Timeline



PASTA Released (LTER)

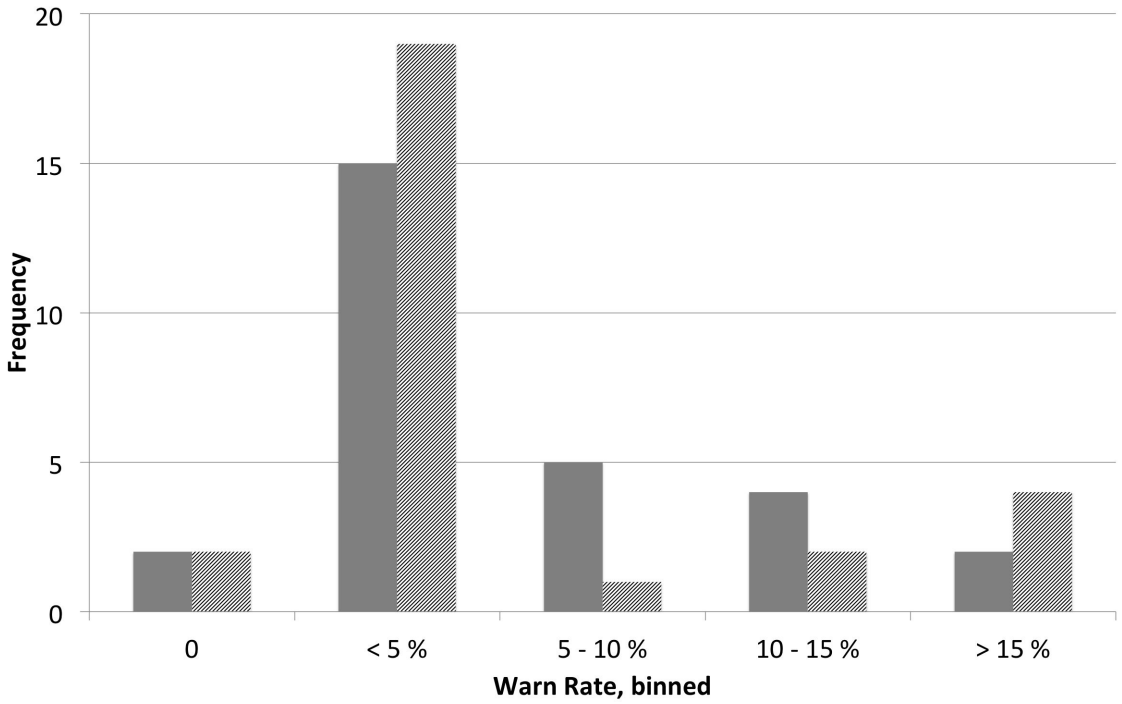
LTER NIS > EDI

Warn Rate by Package Aspect

Dark bars:
Total warn rate (warns/package)

Light bars:
Warn rate for entities only: (warns/entity)

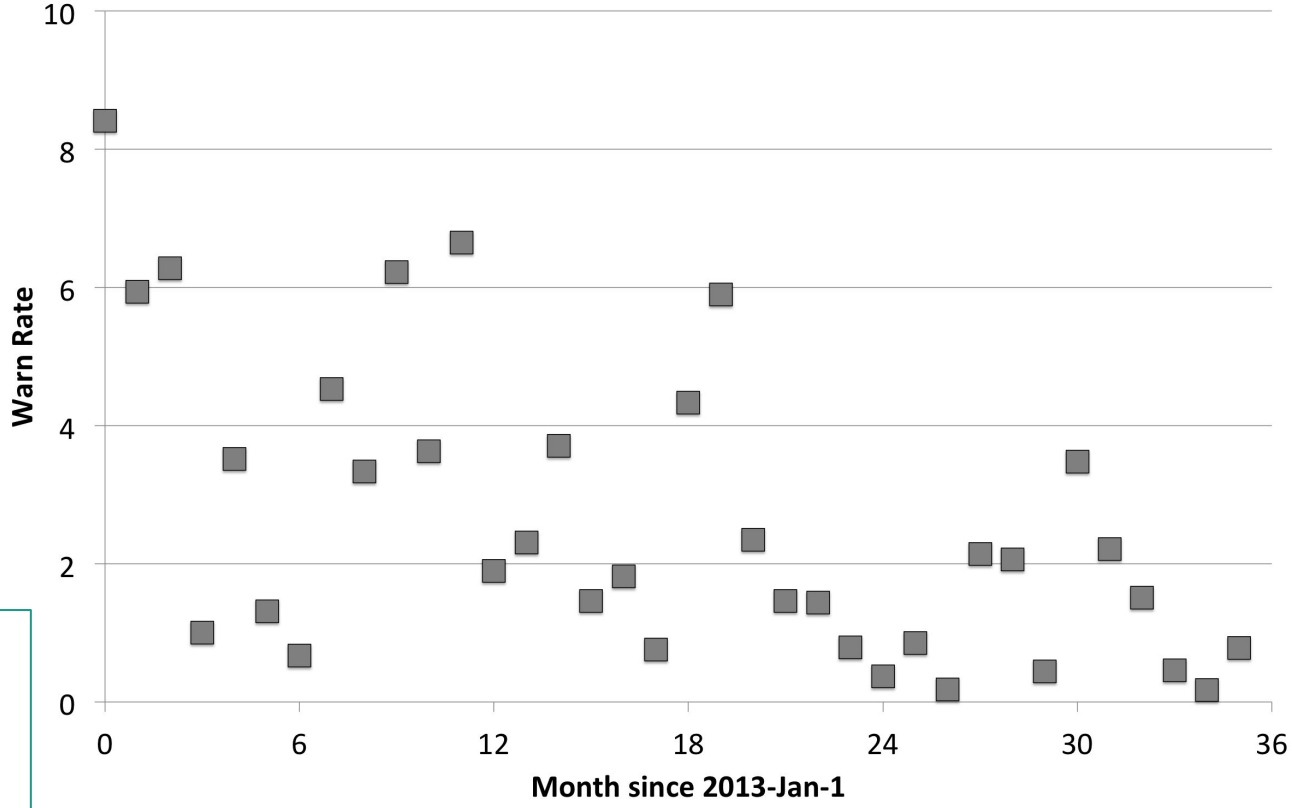
Frequency:
Number of package contributors, 2013-2015



Metadata Evolution

Warn-rate by month*

Decreased from 2013 - 2016



**Warn-rate* = #warns / data package*

Metadata Evolution

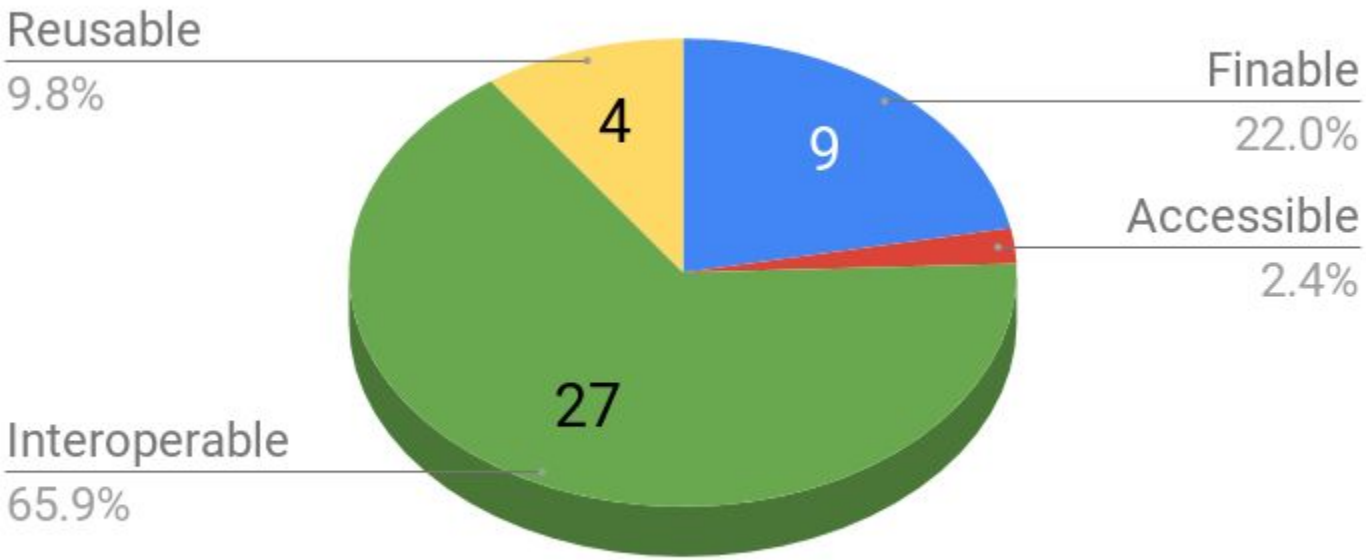
2017:
integrityChecksum
Congruence of entity checksums (found in EML “authentication” node).

fundingPresence:
Presence of project/funding element, to meet requests from NSF

Elements used intermittently (or not at all) until check was implemented; increasing monotonically since

tag	allyear nsites	2013 nsites	2014 nsites	2015 nsites	2016 nsites	2017 nsites	2018 nsites	2019 nsites
abstract	21	18	15	14	23	20	27	27
funding	22	9	9	9	11	13	14	19
maintenance	22	13	10	11	11	13	15	16
personnel	22	10	10	10	13	14	14	19
project	22	10	10	10	13	14	14	19
userId	22	2	1	1	10	10	17	19
additionalInfo	20	8	8	9	8	9	13	15
shortName	20	7	8	7	9	10	14	13
authentication	19					9	15	18

ECC Checks and “FAIR Principles” - Preliminary



Future

Definitely ...

- Support for EML 2.2

Maybe ...

- Align checks with FAIR principles
- Alert-level (between Info and Warn)

O'Brien, M., D. Costa, and M. Servilla 2016. Ensuring the quality of data packages in the LTER network data management system. Ecological Informatics 36: 237–246 DOI: https://doi.org/10.1016/j.ecoinf.2016.08.001

Servilla, M. J. Brunt, D. Costa, J. McGann and R. Waide. 2016. The contribution and reuse of LTER data in the Provenance Aware Synthesis Tracking Architecture (PASTA) data repository. Ecological Informatics 36: 247-258. https://doi.org/10.1016/j.ecoinf.2016.07.003

Take Home - Automated Code Generation = 100%

Because the repository checks data congruence, every data table can be ingested into multiple analysis platforms with scripts that are automatically generated from metadata

https://portal.edirepository.org/nis/mapbrowse?packageid=knb-lter-arc.20067.2

Code Generation: Analyze this data package using: MatLab Python R SAS SPSS tidyr