

EML Congruence Checker

2017-04-10

Margaret O'Brien, Duane Costa

Overview

Name	Description	Response
PastaDoiAbsent	PASTA will add DOIs to L1 EML. No incoming L0 EML should contain a DOI that resembles a PASTA DOI.	error
ChecksumPresent	Checksums can be used to validate entities. inclusion is good practice.	warn
ChecksumMatch	If a checksum is found and does not match the checksum computed by PASTA, package must be rejected.	error
NumRecordsPresent	Check accompanies the current check to compare the number of records in metadata to number found.	warn

PastaDoiAbsent

- PASTA will add DOIs to L1 EML. No incoming L0 EML should contain a DOI that resembles a PASTA DOI.
 - Datasets should have only one PASTA DOI, the one generated and added by PASTA itself, at upload.
 - an 'old' DOI would resolve to the wrong dataset
 - Other DOIs are allowed (e.g., non-PASTA)

Examples:

```
</access>
▼ <dataset scope="document">
  <alternateIdentifier> 268 </alternateIdentifier>
  <alternateIdentifier system="https://doi.org"> doi:10.6073/pasta/321aaa85750bf63250f7fe0f833018d2 </alternateIdentifier>
  <title>
    Nitrate uptake estimates phytoplankton incorporation of dissolved NO3- at selected depths from CCE LTER process cruises
    in the California Current System, 2011 - 2016 (ongoing).
  </title>
  <creator>
```

					attribute's level.
12	pastaDoiAbsent	valid	Type: metadata System: lter On Failure: error	An alternateIdentifier with a DOI system attribute that looks like it is generated by PASTA should not be present	Reject the data package if it contains an alternateIdentifier DOI that looks like PASTA generated it.

Other DOIs

- Other DOIs are allowed (e.g., non-PASTA)

Example

```
</access>
```

```
▼ <dataset>
```

```
  <alternateIdentifier system="https://doi.org"> doi:10.5072/FK2/54d7703c5037abb884baf5c32fba9deb </alternateIdentifier>
```

```
  <title> ME Acadia National Park Lake Nutrient data 1975-2005 </title>
```

```
▼ <creator>
```

```
  ▼ <individualName>
```

```
    <givenName> Bill </givenName>
```

```
    <surName> Gawley </surName>
```

```
  </individualName>
```

Note:

a check has been entered to check that non-PASTA DOIs resolve (not scheduled).

Checksum - 2 checks

- First, check if a checksum has been included in metadata

```
▼<dataTable id="xyz">
  <entityName>NoneSuchBugCount</entityName>
  ▼<entityDescription>
    Bug species richness in the vicinity of Nonesuch, UT
  </entityDescription>
  ▼<physical>
    <objectName>NoneSuchBugCount.txt</objectName>
    <size unit="bytes">882</size>
    <authentication method="SHA-1">915a52bd06ef5730ca5ef33dd359380a59c86ef5</authentication>
    <authentication method="MD5">0c3ff2f519ea20c40a4a15cbdd913d20</authentication>
    <characterEncoding>ASCII</characterEncoding>
  ▼<dataFormat>
    ▼<textFormat>
      <numHeaderLines>0</numHeaderLines>
      <recordDelimiter>\n</recordDelimiter>
      <attributeOrientation>column</attributeOrientation>
      ▼<simpleDelimited>
        <fieldDelimiter>,</fieldDelimiter>
      </simpleDelimited>
    </textFormat>
  </dataFormat>
  ▼<distribution>
```

Checksum - 2 checks

- Second step, compare checksum value to computed value

			On Failure: error			
13	integrityChecksum	valid	Type: congruency System: Iter On Failure: error	Compare the metadata checksum for an entity to the checksum of the downloaded entity	Two possible responses: valid if checksums match; error if checksums do not match.	0c3ff2f519ea2
14	integrityChecksum	valid	Type: congruency System: Iter On Failure: error	Compare the metadata checksum for an entity to the checksum of the downloaded entity	Two possible responses: valid if checksums match; error if checksums do not match.	915a52bd06e
15						

NumRecordsPresent

- Partner to NumRecords
- merely checks that this element is present
- Element is optional, but it is good practice to include it.
 - this check alerts submitters who have not included it
- ```
<attributeList>
 <attribute>...</attribute>
 <attribute>...</attribute>
 <attribute>...</attribute>
 <attribute>...</attribute>
 <attribute>...</attribute>
 <attribute>...</attribute>
 <attribute>...</attribute>
 <attribute>...</attribute>
 <attribute>...</attribute>
 <attribute>...</attribute>
</attributeList>
```
- Example:
- After <attributeList>
- last element of <dataTable>

□

□

[illegible]

# Schedule

- These checks now available on [portal-d.lternet.edu](http://portal-d.lternet.edu):
  - PastaDoiAbsent
  - ChecksumPresent
  - ChecksumMatch
  - NumRecordsPresent
- Deployed to production by end of April 2017
- 
- Up next...
  - Additional checks planned
    - likely ready by end of 2nd quarter
    - warns, only.
-



# Up Next...

Name	Description	type	Response
DateFormatPreferred	Examines <dateFormat> string and compares contents to a list of preferred formats	metadata	warn
DateFormatMatches	Confirms that data values match the format in metadata	congruence	warn
NumericFields	fields stated to be numeric are actually numeric (e.g., not strings)	congruence	warn
Whitespace	examines metadata for certain essential metadata fields for leading/trailing whitespace	metadata	warn

# IMC Input Requested

- New checks need input from the IMC
  - DateFormatPreferred
  - DateFormatMatches
  - NumericFields
  - Whitespace
  -
- Request for an ad hoc working group to determine:
  - List of preferred date formats
  - based on existing datasets and good practice
  -