Infrastructure - Bug #8802

Titles in EML records that use <value> in <title> do not get their titles indexed

2019-05-18 00:55 - Bryce Mecum

Status:	Closed	**Start date:**	2019-05-18
Priority:	Normal	**Due date:**	2019-05-24
Assignee:	Bryce Mecum	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:			
Milestone:	None	**Story Points:**	
Product Version:	*		

Description

Margaret O'Brien over at EDI noticed this and sent it my way. She found an EML record that serializes the title in a schema-valid but unusual way:

```
...snip...
<title>
  <value>Forest-wide bird survey at 183 sample sites the Andrews Experimental Forest from 2009-pre
sent (Reformatted to ecocomDP Design Pattern)</value>
</title>
...snip...
```

See https://search.dataone.org/view/https://pasta.lternet.edu/package/metadata/eml/edi/359/1 and notice the citation is missing its title (which is powered by Solr) and also see the accompanying Solr doc at https://search.dataone.org/cn/v2/query/solr/?q=id:%22https://pasta.lternet.edu/package/metadata/eml/edi/359/1%22 and notice the title field is not set.

This is a bit tricky. It's pretty clearly *not* endorsed in the EML spec, as per:

i18nNonEmptyStringType: (emphasis mine)

> This type specifies a content pattern for all elements that require language translations. The xml:lang attribute can be used to define the default language for element content. *Additional translations should be included as child 'value' elements* that also have an optional xml:lang

So value in title is intended to be used like this:

```
<title>
  My title
  <value xml:lang="fr">Mon titre</value>
</title>
```

and not how it is in https://search.dataone.org/view/https://pasta.lternet.edu/package/metadata/eml/edi/359/1.

Do we want to tweak the indexer to support this or is it actually a good thing that the indexer didn't pick this up because it's, subjectively, not well-formed EML?

History

#1 - 2019-05-18 00:57 - Bryce Mecum

Also, the eml-base bean defines the EML title field as:

```
<bean id="eml.title" class="org.dataone.cn.indexer.parser.SolrField">
  <constructor-arg name="name" value="title"/>
  <constructor-arg name="xpath" value="//dataset/title/text()"/>
  <property name="multivalue" value="false"/>
</bean>
```

which explains why the title being stored in element data wasn't picked up.

#2 - 2019-06-06 19:39 - Bryce Mecum

- % Done changed from 0 to 100

- Status changed from New to Closed

We talked about this on a dev call the other week and decided that the added complexity to cover this case wasn't worth it, especially given how rare we expect this to be. The EML docs themselves even indicate this pattern (value inside title) isn't the intent of the type (i18nNonEmptyStringType):

> Additional translations should be included as child 'value' elements that also have an optional xml:lang attribute.

I messaged Margaret O'Brien to hopefully communicate this decision back to the content producer in question and get the EML record updated.