EML - Bug #3232

EML parser limitations

04/17/2008 11:32 AM - Margaret O'Brien

Status:	Closed	**Start date:**	04/17/2008
Priority:	Immediate	**Due date:**	
Assignee:	Matt Jones	**% Done:**	0%
Category:	eml-parser	**Estimated time:**	0.00 hour
Target version:	EML2.1.0	**Spent time:**	0.00 hour
Bugzilla-Id:	3232		

Description

This is just for the record. It seems that the EML parser could benefit from an update, although it's current behavior is perfectly legal.

It may be that bug 2054 appeared because the parser that comes with EML does not use schema-full-checking. My main resourse (Walmsley 2002 book) says that this is the xerces feature that checks for non-deterministic content models (which was the error in 2054). That feature doesn't appear to be in the file SAXValidate.java - at least not to my untrained eye.

Bug 2703 seems to have come about because Xerces does not necessarily load all the import schemas. The content model for appinfo and documentation is a wildcard, and can be validated laxly. So it's up to the validator to go looking for element declarations, but it doesnt have to. This behavior is perfectly legal.

So the parser can detect errors instance documents, but it does not adequately catch schema errors. Maybe this was always the intent, but not quite clearly stated. Or, maybe it's a simple matter to add some other xerces features, or incorporate XSV instead - but not being a java programmer, I dont know.

History

#1 - 04/17/2008 11:41 AM - Matt Jones

If the EML Parser can be made to check additional schema validity constraints, we should go ahead and fix this for the 2.1.0 release. Any lack of checking on the schema validity in previous releases was an error.

The EML Parser is the definitive check on EML validity, and is used in many applications, such as Metacat and Morpho. One difficult side effect of upgrading the parser would be that Metacat would no longer accept the documents that it accepts today (i.e., 2.0.1 docs) if the schemas don't validate, and so this could trigger a large set of problems for content providers from various sites by forcing an upgrade to 2.1.0.

#2 - 04/22/2008 10:49 AM - Margaret O'Brien

I'm concerned that a requirement to rewrite the parser first will significantly impede the release of 2.1.0. Obviously, applications like metacat can't reject all the 2.0.x documents, and so the solution will be complicated.

Is it reasonable to split the parser from EML? or to consider it only an instance-parser in 2.1 (since it does this adequately), and put the new parser in a future release? During schema editing, a parser included with EML would be useful to someone who did not have access to another one. But given that free schema editors are now quite good at that (albeit not perfect), it may not be absolutely necessary to include it.

#3 - 04/23/2008 12:17 PM - Callie Bowdish

These are two tests were to see if the older versions of eml.2.0.1 would validate with the newer version. Very simple data packages validated but these two did not. Very basic data packages did validate.

This is an eml.2.1.0rc3 validation error using xerces version -2_9_1.1. These data packages have not been edited to conform to new eml requirement for describes metadata.

[Error] 876.3:61:32: cvc-complex-type.2.4.a: Invalid content was found starting with element 'unitList'. One of '{describes, metadata}' is expected.
876.3: 1290 ms (66 elems, 34 attrs, 0 spaces, 345 chars)

eml validation with with test document at http://dev.nceas.ucsb.edu/knb/metacat?action=read&qformat=knb&docid=bowdish.876.3

This is data package created with 2.0.1 eml editor only the eml version sections were changed at the top of the document

[Error] 251.9:47:59: cvc-type.3.1.1: Element 'references' is a simple type, so it cannot have attributes, excepting those whose namespace name is identical to 'http://www.w3.org/2001/XMLSchema-instance' and whose [local name] is one of 'type', 'nil', 'schemaLocation' or 'noNamespaceSchemaLocation'. However, the attribute, 'system' was found.
251.9: 1337 ms (221 elems, 45 attrs, 0 spaces, 9500 chars)

http://dev.nceas.ucsb.edu/knb/metacat?action=read&qformat=knb&docid=bowdish.251.9

---
eml 2.0.1 validates on these two test data packages
this section is edited to test 2.0.1 <?xml version="1.0"?><eml:eml xmlns:eml="eml://ecoinformatics.org/eml-2.0.1" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" packageId="bowdish.251.9" scope="system" system="knb" xsi:schemaLocation="eml://ecoinformatics.org/eml-2.0.1 http://knb.ecoinformatics.org/knb/schema/eml-2.0.1/eml.xsd">

This section is edited to test 2.1.0rc3

<?xml version="1.0"?><eml:eml xmlns:eml="eml://ecoinformatics.org/eml-2.1.0rc3" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" packageId="bowdish.251.9" scope="system" system="knb" xsi:schemaLocation="eml://ecoinformatics.org/eml-2.1.0rc3 eml.xsd">

#4 - 04/23/2008 12:24 PM - Matt Jones

Regarding Comment #2 about not needeeding the EML Parser, I disagree. A valid EML document requires more than just schema validity, as there are several validity rules spelled out in the specification that must be enforced but are not part of the XSD schema. The EML Parser currently checks these, and so I think it is a mistake to release a version of EML without a properly functioning EML Parser. It has been a part of all prior releases.

#5 - 04/23/2008 01:27 PM - Margaret O'Brien

With regard to comment #3 (2 failed docs):
1. the first test shows how many 201 docs will fare when validated against 210 -- the now-required <metadata> element is missing. The only 201 docs which will pass will be those from people who pre-implemented the bug 2054 fix.

2. the second error is more complicated. There had been a partial fix of bug 1662 that was checked into the head, but was not in the download zip of eml201. See comment #4 on bug 1662: the code which added attributes to <references> was merged into a branch with other target-unspecified changes. Since it had been in the head, there was always the possibility that someone with a recent checkout might use it (as Callie has). If the use of that attribute is pervasive, and if it's acceptable to address 1662 partially for now, then it can go back in the head. advice/opinions, please.

#6 - 07/10/2008 12:01 PM - Margaret O'Brien

Hi Jing -
examples from the Walmsley 2002 book are online, but the text is not. My print copy (Ch 5, p91-92) says that schema-full-checking is the feature that checks for non-deterministic content models, which is what bug 2054 revealed about EML2.0.1

See
http://www.datypic.com/books/defxmlschema/chapter05.html
Here's walmsley's example code:

Example 5-8. Java code to set schema validation in Xerces

```
SAXParser p=new SAXParser();
try {

// Turn schema validation on
p.setFeature
("http://xml.org/sax/features/validation", true);
p.setFeature
("http://apache.org/xml/features/validation/schema", true);
p.setFeature
("http://apache.org/xml/features/validation/schema-full-checking",
true);
} catch (SAXException e) {
System.out.println("error turning on validation");
}
```

whereas in eml/src/org/ecoinformatics/eml/SAXValidate.java it's like this:

```
...
if (schemavalidate) {
parser.setFeature(
"http://apache.org/xml/features/validation/schema",
true);
}
...
```

which Walmsley says "allows most schema checking to take place"

BTW, googling for schema-full-checking xerces turned up this index on both the xerces page and stylus studio (xmlSpy) so maybe that means that spy is using xerces?
http://xerces.apache.org/xerces2-j/javadocs/xerces2/index-all.html
http://www.stylusstudio.com/api/xerces2/index-all.htm

#7 - 07/10/2008 12:52 PM - Jing Tao

Hi, Margaret: here is what i read online from xerces site:

http://apache.org/xml/features/validation/schema-full-checking

True: Enable full schema grammar constraint checking, including checking which may be time-consuming or memory intensive. Currently, unique particle attribution constraint checking and particle derivation restriction checking are controlled by this option.
False: Disable full constraint checking.
Default: false
Note: This feature checks the Schema grammar itself for additional errors that are time-consuming or memory intensive. It does not affect the level of checking performed on document instances that use Schema grammars.

I am not sure if xmlspy uses xerces or not :(

#8 - 07/10/2008 04:09 PM - Jing Tao

After adding this statement at SAXValidate.java:
parser.setFeature("http://apache.org/xml/features/validation/schema-full-checking",
true);
it seems that we can valid eml schemas themselves now (it rejected an eml201 document). However, when we tried to valid eml210rc3 document, it seems stmml.xsd has some problem:
cos-nonambig: ("http://www.xml-cml.org/schema/stmml":definition){0-1} and ("http://www.xml-cml.org/schema/stmml":definition) (or elements from their substitution group) violate "Unique Particle Attribution".

#9 - 07/10/2008 04:19 PM - Jing Tao

In EML module, there are two parsers:
EMLParser.java and SAXValidate.java. EMLParser will check the unique id and reference id stuff (only for eml features) and SAXValidate will valid EML instance against schema and valid EML schema itself.

In metacat, we use EMLParser part, not SAXValidate. So this change wouldn't affect metacat at all.

In morpho, it has a similar class to SAXValidate of EML module. It doesn't use either EMLParser or SAXValidate from EML module. So this change wouldn't affect morpho either.

#10 - 03/27/2013 02:22 PM - Redmine Admin

Original Bugzilla ID was 3232

#11 - 07/03/2014 11:16 AM - Matt Jones

- Target version changed from EML2.1.0 to EML2.2.0

Need to check if this has now been fixed, and if so, close the bug. Margaret, Jing, any thoughts?

#12 - 07/03/2014 03:51 PM - Jing Tao

- Status changed from In Progress to Closed

- Target version changed from EML2.2.0 to EML2.1.0

I just checked and found it was fixed in the 2.1.0 release.

The solutions is that: eml document will be checked by setting "http://apache.org/xml/features/validation/schema-full-checking" true. But the exceptions is set for the eml 2.0.0 and eml 2.0.1. Because of the issues in eml 2.0.0 and 2.0.1 schemas, the eml document will be rejected if the "http://apache.org/xml/features/validation/schema-full-checking" is true even though the document is legal.