EML - Bug #586

resolve validation problem with missing keys

09/10/2002 10:17 AM - Matt Jones

Status:	Resolved	**Start date:**	09/10/2002
Priority:	Immediate	**Due date:**	
Assignee:	Matt Jones	**% Done:**	0%
Category:	eml - general bugs	**Estimated time:**	0.00 hour
Target version:	EML2.0.0rc2	**Spent time:**	0.00 hour
Bugzilla-Id:	586		

Description

The new xerces reports a validation error for the "eml" module using ant test.
This seems to be because keys are defined that are not used. Nedd to determine
if this is truly invalid or a xerces bug. If its invalid, we'll need to change
our key system becuase it won't work with optional key fields.

History

#1 - 09/10/2002 10:20 AM - Peter McCartney

I think this is also the reason why we're having problems with Stylus Studio
which is based on xerces. It reports a flood of errors related to key
statements.

#2 - 09/11/2002 11:51 AM - Matt Jones

OK, here's the deal as far as I can tell with the key validity problems. These
surfaced when we upgraded Xerces, which became more sensitive to these schema
validity problems. When you define a key, you define a set of nodes to be
checked using the "selector" xpath expression. This produces a node set, and
for each field in that node set, it checks that the "field" selector contains a
unique value AND is not nil.

This means two things for us:
1) id attributes can not be optional, because we have to have a value to pass
the validation test
2) our selectors can't look at every node (as they currently do by using
something like ".//*"), because such a broad selector will return nodes that
don't allow the id attribute, and thus would not have a value for the key

The only solution that I can see right now is:
1) change all id attributes from optional to required
2) change our xpath selectors to only select eml fields that contain an id attribute

This is a pretty major change, but as it stands now the key/keyref stuff
prevents us from ever creating a valid document. Comments appreciated.

#3 - 09/11/2002 12:23 PM - Chad Berkley

I think the solution you propose sounds simple enough. I don't see required id
fields as being a big issue, although I know it has been an issue with others
before. The xpath selection seems trivial. We definitely need this stuff to
validate correctly with existing tools or we may have a mess on our hands.

#4 - 09/12/2002 04:19 PM - Peter McCartney

needless to say, im one of the ones that is not wild about id's being
required. when we agreed on the identifier/reference thing, it was with the
understanding that one could produce an eml with duplicated content or choose
to use identifiers.

I think eml should be about content, not schemas. to me, all this keyref stuff
is about enforcing one opinion on the mangement of similar information within
one document - it has nothing to do with content standards. If we're facing a
problem caused by key refs that is going to force everyone to have to create
documents following this narrower opinion, I would sooner leave them optional
and let external software do the enforcing. In order to used them effectively,
youre dependent on software to manage them anyways.

#5 - 09/13/2002 05:59 AM - Tim Bergsma

Now I'm scared. I've been largely ignoring the identifier discussion, probably because I thought they wouldn't affect me. When triples were dropped, I suddenly felt that I might be able to generate eml independently, without expert intervention. Peter's comments about identifiers are unsettling. I'm sure we'll get it figured out, but at what cost, on a site basis? I'll be paying more attention now.

#6 - 09/13/2002 07:23 AM - Chad Berkley

Alas, all is not lost! Matt and I spent a bunch of time yesterday tracking down exactly what was going wrong with the keys. We feel pretty strongly that this is a bug in xerces and that we are doing it correctly. Matt entered a bug in the xerces bugzilla: http://nagoya.apache.org/bugzilla/show_bug.cgi?id=12592 If you look at it, you can see what we figured out as well as some sample documents that illustrate our points.

#7 - 09/13/2002 03:54 PM - Matt Jones

Actually, the crux of the issue still remains in that xerces seems to be properly throwing an error if elements that particpate in a key are missing an id. The seem to be saying that all keys have the "NOT NULL" property. Bummer. It means there is nothing like the "ID" type from XML 1.0 in XML Schema. And this is why id would need to be required. I'm not at all satisfied with the "leave it optional and let people submit invalid docs", because that will never work. So...I think we need to continue this duscussion post RC1. In the meantime, I've commented out the key and keyref elements from eml.xsd so that everything validates. I'll update the README to indicate this is an outstanding problem.

The xerces folks are working on the othe rhalf of the problem, but its less critical as there would be a work-around.

#8 - 10/04/2002 10:43 AM - Matt Jones

OK, so here's the deal. We've agreed ID should not be required. So that rules out using XML Schema keys to force uniqueness of the ids, because they do not allow nulls (if you select an element for a uniqieness check, its an error if it doesn't have the id at all). So...Chad wrote an EMLParser in Java that does the checks they way we want them done. It basically takes the key/keyref defs and validates that they follow the rules laid out in the eml spec. This code is shipping with the EML release, and has been exposed as a servlet on the web for people to use for validating their eml instances.

Hopefully Xerces will fix their bug, but the more serious problem would still remain, so our custom parser seems like the best solution for now. Will need to flag this in the distribution.

RESOLVED FIXED.

#9 - 03/27/2013 02:14 PM - Redmine Admin

Original Bugzilla ID was 586