EML - Bug #1662

id key definitions in EML

08/26/2004 05:24 PM - Peter McCartney

Status:	In Progress	**Start date:**	08/26/2004
Priority:	Immediate	**Due date:**	
Assignee:	Matt Jones	**% Done:**	0%
Category:	eml - general bugs	**Estimated time:**	0.00 hour
Target version:	EML2.2.0	**Spent time:**	0.00 hour
Bugzilla-Id:	1662		

Description

there are several problems emerging with the unique key definitions in eml. in eml.xsd, there is a key definition that requires all instances of the @id attribute to be unique within the document. when content is to be duplicated there is to be one instance of the content with an id assigned and all other instances are to use the <references> tag to point to that id. id's in a document may be declared as document or system scope, meaning that they are declared to be unique only within the document or within a broader naming authority that is identified in the @system attribute.

Here are some of the problems:
1. there no enforced unique constraint on the @system attribute, although it is implicit. Thus it is possible to create a dataset element usint an id with system=cesdataset and a creator with system=asupersonnel. when those systems have each assigned a similar value, you get conflicts. to avoid them, users are forced to change identifiers and break the pointers back to the original source of the content.

2. the spirit of the id and references tags was to insert some degree of normalization that xml inherently lacks. however, it can be a rather arbitrary choice with in the document which instance is the one that gets the content and which ones get the reference pointer. This makes it very difficult for people trying to write tools to edit eml documents since one could easily drop an element that contains elements that contain content that other elements are pointing to. This gratuitously complicates programming for EML and is likely to discourage potential contributors of tools for working with the standard.

3. EML method allows you to embed the eml of related datasets that were used to produce the current one in the methods discussion. conflicts can arise between the identifiers of the embedded datasets. attemping to resolve conflicts between documents using references could mean you have to edit those documents rather than embed them.

History

#1 - 09/01/2004 02:34 PM - Peter McCartney

I checked in changes to eml-resource.xsd and buildDocbook.xsl as a proposed fix to some of this bug. The fix introduces an optional system attribute to the references element, which corresponds to the existing system attribute accompanying the id element through out EML. if used, it qualifys that the matching id pointed to by the references tag has a scope defined by the system attribute. Thus, EML documents can contain more than one id attribute with the same value provided that they are scoped with different system attributes. Changes to the "reusable content" section reflect this. if system is defined, it must be defined for both the target and the referencing element and it must match. Otherwise it must be absent from both.

I think there is still some ambiguity in the wording of the docs that could benefit from some additional input. the eml root has an attribute called scope which is "document" by default and "system" by option. We now allow there to be more than one system referenced. if system tags are used once, does this mean they must be provided in all cases? or, can i write an eml where the ids for attribute are document scope, but the ids supplied for literature cited are

scoped to a particular system? Do we really need the "scope" attribute? if it is left as default, is it invalid to use the system atrribute anywhere in the document?

#2 - 09/02/2004 09:34 AM - Matt Jones

Changing target milestone for this feature.

#3 - 09/02/2004 09:38 AM - Matt Jones

Changing QA contact to the list for all current EML bugs so that people can track what is happening.

#4 - 04/08/2008 02:36 PM - Margaret O'Brien

The consensus on this bug seems to be that it will require more testing before implementation can proceed, and that it is not appropriate for 2.1.0. So the r1.78 has replaced it in the head. I also merged the changes that Peter committed (in r1.79) into the branch which is currently called eml_2_1_0.

So now there's a problem; this branch name needs to change for obvious reasons -- eml2.1.0 is on the trunk. The branch contains fixes which seem to now be mostly target-unspecified: 1132 (access control issues), 1152 (an element name change that may be dropped), and at least one (#1031) that appears to be already in 2.0.1. Retagging it with any version number seems confusing until a plan for future versions has evolved. Should this branch be renamed eml_target_unspecified, and its contents re-evaluated?

#5 - 10/01/2008 05:02 PM - Matt Jones

After reviewing this bug, I still concur that it is an important problem that should be addressed as it causes validation problems for people trying to link to external systems. However, the exact implementation will require a more sophisticated EML Parser that can check if IDs are unique within the context of a particular system attribute. So, I think this additional work requires that we move this fix to the next release so that EML 2.1 can be released without delay. I have retargeted this for EML 2.2.0, and it should be triaged then when we make plans for that release.

One other note: In his original report, McCartney worries that it will be difficult to build applications that handle references in arbitrary places. With experience doing this for Metacat and Morpho, it now dow not seem to be such a big problem, so I think that is a non-issue.

#8 - 04/15/2011 12:41 PM - gastil gastil

Prompted by a discussion about id, scope, and system in the lno-nis, I asked Margaret for clarification of this issue. Attached are two example EML docs, one using the current 2.1.0 and one that may work when this feature is added perhaps in 2.2.0

In EML 2.1.0 the way it is now,
some elements can have none or all of the attributes
id
scope
system

And the value of the scope attribute can be either
document
system

Currently, the EML parser requires that the value of the id attribute be unique within an EML document, irregardless of the presence and values of the scope and system attributes.

The feature request in bug 1662 asks that in EML 2.2.0 that the combination of (id, scope, system) be unique, rather than requiring id alone to be unique.

It would still require that id values with the scope of document be unique. There could not be two identical id values both with scope document.

The "example_id_system_scope.xml" (attached and also available as http://dev.nceas.ucsb.edu/knb/metacat/knb-lter-xxx.1) parses. The "example_id_system_scope_future.xml" does not parse in EML 2.1.0 but we suggest it would parse if this feature is added. The same id, janedoe, is used twice, but with different scopes. The use of that id in a reference raises the issue of how that reference would be disambiguated.

Is that clarification accurate?

#9 - 01/10/2012 02:24 PM - Margaret O'Brien

This would also mean that only ids with the scope="document" could be used in <references>, since ids with scope="systeM" might not be unique, and unique ids for referencing is very reasonable.

So in the example above (id_system_scope_future.xml),

this id could be used in a reference:
<creator id="foo" scope="document">

but this one could not:
<contact id="janedoe" scope="system" system="LTER">

Another task for the parser.

#10 - 03/27/2013 02:17 PM - Redmine Admin

Original Bugzilla ID was 1662

Files

example_id_system_scope.xml	1.89 KB	04/15/2011	gastil gastil
example_id_system_scope_future.xml	2.13 KB	04/15/2011	gastil gastil