EML - Bug #602

eml-physical

09/24/2002 12:23 PM - Owen Eddins

Status:	Resolved	**Start date:**	09/24/2002
Priority:	Immediate	**Due date:**	
Assignee:	Owen Eddins	**% Done:**	0%
Category:	eml - general bugs	**Estimated time:**	0.00 hour
Target version:	EML2.0.0rc2	**Spent time:**	0.00 hour
Bugzilla-Id:	602		

Description

Matt Jones pointed out that asciiFixed and asciiDelimited should
be changed to less misleading name like textFixed and textDelimited
because other encoding schemes are possible. Unicode for example.
Describing these as asciiFixed or asciiDelimited is misleading
because it implies it can only be ascii. The encoding scheme
can be set in <physical><dataObject><characterEncoding>

Data Objects whose format is a mixture of fixed and delimited are
not supported as eml-physical is currently structured. For example,
data objects whose physical structure looks like this cannot be
represented.

May,100aaaa,1.2,
April,200aaaa,3.4,
June,300bbbb,4.6,

The second attribute is a composite of two attributes that
are of fixed length but with no fixed fieldStartColumn.

I recommend the following changes to eml-physical to support mixed
data formats.

Both asciiDelimited and textFixed be placed as repeatable choices
under a new element called textFormat. numHeaderLines and
numPhysicalLines be made optional subelements of textFormat
because they are global to the data objects being described.
The only actual content change would be moving numPhysicalLines
as a subelement of textFixed and making it a subelement of
textFormat. So the instance document chunk that would describe
the above data object would look like this:

<physical>
<dataObject>
.
.
.
</dataObject>
<dataFormat>
<textFormat>
<textDelimited>
<fieldDelimiter>,</fieldDelimiter>
</textDelimited>
<textFixed>
<fieldBounds>
<fieldStartColumn>-1</fieldStartColunm>
<fieldWidth>3<fieldWidth>
</fieldBounds>
<fieldBounds>
<fieldStartColumn>-1</fieldStartColunm>
<fieldWidth>4<fieldWidth>
</fieldBounds>

</textFixed>
<textFixed>
</textFixed>
<textDelimited>
<fieldDelimiter>,</fieldDelimiter>
</textDelimited>
</textFormat>
</dataFormat>
</physical>

Note that fieldStartColumn is set to -1. Because this column does not make sense in a mixed format context we could set this value to -1 OR make this element optional. Currently fieldStartColumn is an unsignedInt. We would have to make it an integer or long to support negative numbers

See file eml-physical.xsd sent to eml-dev.

The above solution for mixed data formats solves the problem of <asciiFixed><fieldBounds> not being repeatable. If folks want eml-physical to stay as is this element needs to be made repeatable or you will be limited to only one attribute per dataObject. Clearly, this was an oversight.

History

#1 - 09/24/2002 12:25 PM - Owen Eddins

I'm attaching eml-physical.xsd instead of sending in to eml-dev

#3 - 10/01/2002 05:18 PM - Matt Jones

Reformatted and checked in owen's schema as attached here. Some notes and issues:

1) Many of the fields are now only available under the "complex" element, when in fact they would be useful generally in both the simpleDelimited and complex cases. It would be useful to factor out the recordDelimiter, quoteCharacter, literalCharacter fields for use in both delimited cases. Also, maxRecordLength really applies to both fixed and delimited cases, and is just a guide to tell some processing systems what a physical record might look like. There are cases in fixed format data where there is no recordDelimiter, and one must rely on maxRecordLength to determine when to stop parsing one record and start parsing the next. Sooo... is it ok if I move the location of these fields?

2) I don't understand how lineNumber works in the current model of complex. Could you explain how I should interpret a line number of "2" in the current model? Maybe an example instance of a multi-line record would be useful in a test/multiline.xml example. We already have samples of the others in test/eml-physical.xml and lib/samples/eml-sample.xml for reference -- they validate properly now, but will need to be updated for any changes made.

3) Also, a bunch of fields are not documented, and some of the documentation seems out-of-date -- we need to make sure that this documentation is complete and accurate, although I can understand waiting until we finalize the schema.

#5 - 10/03/2002 11:25 AM - David Blankman

If orientation moves from dataTable to physical then it should probably be modeled differently in binaryRaster, textFormat, and formatType. In binaryRaster it should have an enumeration of "landscape" or "portrait". In textFormat, "recordsInRows" or "recordsInColumns". Maybe we should rename formatType to otherFormatType. In other FormatType, perhaps a union of "recordsInRows" or "recordsInColumns" and any.

#6 - 10/04/2002 10:07 AM - Matt Jones

Finished schema changes to eml-physical, and completely redocumented the module. Moved orientation from eml-dataTable to eml-physical because, after much discussion, we agreed this is a physical construct. As far as I know, all changes to eml-physical for RC2 are complete. RESOLVED FIXED.

#7 - 03/27/2013 02:14 PM - Redmine Admin

Original Bugzilla ID was 602

Files

eml-physical.xsd	69.1 KB	09/24/2002	Owen Eddins
eml-physical.xsd	68.7 KB	10/01/2002	Owen Eddins
eml-physical.xsd	71 KB	10/02/2002	Owen Eddins