EML - Bug #585

internationalization needed in EML

09/06/2002 03:56 PM - Chad Berkley

Status:	Closed	**Start date:**	09/06/2002
Priority:	Low	**Due date:**	
Assignee:	Matt Jones	**% Done:**	0%
Category:	eml - general bugs	**Estimated time:**	0.00 hour
Target version:	EML2.1.1	**Spent time:**	0.00 hour
Bugzilla-Id:	585		

Description

We need ot investigate possible internationalizing EML. Maybe by using enabling the use of 'lang' attributes.

History

#1 - 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.

#2 - 12/05/2008 09:31 AM - Margaret O'Brien

This comment from an email from David Blankman:
As EML is becoming an international standard, we need to start thinking about ways to make EML more intelligent about multiple languages. While EML allows multiple titles, there is currently no way to indicated that multiple titles are equivalent. For example,if I have:
<title> North American Forests </title> AND
<title> Bosques de Norte Americano</title>

EML currently has no way to indicate that these are the same title, just in a different language.

Matt and I were talking about this at the ISEI-Cancun meeting, but I thought that it would be a good idea to get this discussion started within eml-dev and the ILTER group as well.

#3 - 12/08/2008 10:33 AM - Matt Jones

David and I discussed (briefly) some of these issues at ISEI. And we also discussed them at the ILTER meeting in China. The 'language' tag in eml-resource defines the language of the resource, which in the case of eml-dataset resources means the language of the data. Interestingly, we don't really have a language tag per se for the EMl document itself, except that all XML documents can use the built-in "xml:lang" attribute, which is optional for all XML elements (http://www.w3.org/TR/REC-xml/#sec-lang-tag). This allows one to set the language for each and every element in an XML document, such as:

<title xml:lang="en">North American Forests</title>
<title xml:lang="es">Bosques de Norte Americano</title>

Two problems we would need to address with this approach come immedately to mind:

1) Many elements in EML are not repeatable, and therefore it is not possible to have one copy of the element in English and another in a different language. So cardinality would have to be updated throughout the EML schemas, which would make some aspects of validation more confusing.
2) For those elements that are already repeatable or are made repatable through a revision, there is no mechanism to indicate that the two element nodes are meant to be have the same semantic meaning in different languages, as opposed to two semantically different elements that happen to also differ in their language.

This second issue is the one that would require more structural changes to EML. For example, one might sometimes want to have more than one title (which is why title is currently repeatable), but other times want to have one title in two different languages. Either way, EML's current structures don't allow these subtleties to be specified.

Matt

#4 - 12/09/2008 03:15 PM - Matt Jones

After more conversation on the email list, it seems that the approach to localization used in ISO 19139 may be applicable here. A fragment of a 19139 document might include:

<abstract xsi:type="PT_FreeText_PropertyType">
<gco:CharacterString>Brief narrative summary of the content of the resource</gco:CharacterString>
<!--== Alternative value ==-->
<PT_FreeText>

<textGroup>
<LocalisedCharacterString locale="#locale-fr">RÃ©sumÃ© succinct
du contenu de la ressource</LocalisedCharacterString>
</textGroup>
</PT_FreeText>
</abstract>

So, the PT_FreeText_PropertyType is very similar in concept to the EML TextType. We could indeed define a new LocalizedTextType that use this same trick, basically allowing textGroup subelements with alternate language strings. Or we could simply use the definition of PT_FreeText in EML via an import (except that there may be restrictions on free reuse of ISO standards, which would prevent us from incorporating such a thing directly in EML, as redistribution is critical to an open standard). Although the naming conventions they've used are not particularly readable.

Note that this approach depends on a previously defined locale (locale="#locale-fr"), which is provided by a different set of elements earlier in the metadata for defining locales. The local defines both a language code and a character encoding for strings in that locale:

```
&lt;locale&gt;
    &lt;PT_Locale id="locale-fr"&gt;
      &lt;languageCode&gt;
        &lt;LanguageCode
           codeList="resources/Codelist/gmxcodelists.xml#
LanguageCode"
           codeListValue="fra"> French &lt;/LanguageCode&gt;
      &lt;/languageCode&gt;
      &lt;characterEncoding&gt;
        &lt;MD_CharacterSetCode
         codeList="resources/Codelist/gmxcodelists.xml#
MD_CharacterSetCode"
         codeListValue="utf8">UTF 8&lt;/MD_CharacterSetCode&gt;
      &lt;/characterEncoding&gt;
   &lt;/PT_Locale&gt;
  &lt;/locale&gt;
```

This is powerful, but it seems to me that the characterEncoding could get one in trouble with XML parsers if the locale character encoding differs from the encoding defined in the XML Prolog. As far as I know, an XML document can have one and only one character encoding, so mixing different elements with different encodings will probably mess up standard XML processors. This would have to be explored to see if it is a significant issue.

#5 - 07/21/2010 04:41 PM - ben leinfelder

Looking into this more concretely, it doesn't appear to be overly daunting.
For the dataset/title element I've extended the non-empty string type that it currently is in EML 2.1.0 to allow "mixed" content so that the original title can remain unchanged and additional translations can be added. This means multi-language fields would not need to be modified en masse when upgrading documents from existing EML 2.1.0 to EML 2.1.1 (or whatever EML version number we choose).
Fragment supporting old and new:

<title>
Original title
<!-- language translations -->
<value xml:lang="en">Title in English</value>
<value xml:lang="es">Titulo en Español</value>
</title>

The relevant XSD change is below:

<xs:complexType name="i18nNonEmptyStringType" mixed="true">
<xs:sequence>
<xs:element name="value" minOccurs="0" maxOccurs="unbounded">
<xs:complexType>
<xs:simpleContent>
<xs:extension base="NonEmptyStringType">
<xs:attribute ref="xml:lang" />
</xs:extension>
</xs:simpleContent>
</xs:complexType>
</xs:element>
</xs:sequence>
</xs:complexType>

While this is valid XML that conforms to a valid schema, I'm not sure the "mixed" type will be the easiest to work with. Presumably XML parsers can deftly handle mixed elements, but I imagine there will be a few unanticipated gotchyas. Certainly for xPath-like queries, we'd need to be explicit about searching both the original element and any possible localized versions (sub-elements) of it, but this is an inescapable hurdle for any internationalization solution.
The current 'DocBook' style structure for some text fields (the "abstract" comes to mind) already uses "mixed" elements where text and structure can be interleaved. I believe multi-lingual extensions to those elements would be similarly straight forward as what I've described above.

#6 - 03/27/2013 02:14 PM - Redmine Admin

Original Bugzilla ID was 585

#7 - 04/16/2014 01:56 PM - Matt Jones

- Status changed from In Progress to Closed

- Target version changed from EML2.2.0 to EML2.1.1

The i18nNonEmptyStringType described above was implemented and released as part of EML 2.1.1 -- closing this ticket as the feature is now complete.