EML - Bug #1197

dictionary needed for externallyDefinedFormat

10/31/2003 09:17 AM - Peter McCartney

Status:	New	**Start date:**	10/31/2003
Priority:	Normal	**Due date:**	
Assignee:	Matt Jones	**% Done:**	0%
Category:	eml - general bugs	**Estimated time:**	0.00 hour
Target version:	Postpone	**Spent time:**	0.00 hour
Bugzilla-Id:	1197		

Description

Externally defined format is useless for automatic processing unless you have some idea what to look for. This is a step backwards from FGDC which at least provided enumerations for the common file formats at the time.

History

#1 - 12/17/2003 10:36 AM - Peter McCartney

Here is a possible format for a dictionary file to provide an anuthority and reference for data formats (and archive formats)

#2 - 12/17/2003 12:19 PM - Peter McCartney

```
<?xml version="1.0" encoding="UTF-8"?>
<dataFormats>
<externallyDefinedFormat name="Shapefile" description="ESRI shapefile" >
<part extension="shp" mime="application/octet-stream"/>
<part extension="dbf" mime="application/octet-stream"/>
<part extension="shx" mime="application/octet-stream"/>
<part extension="prj" mime="application/octet-stream"/>
<part extension="idx" mime="application/octet-stream"/>
</externallyDefinedFormat>
<externallyDefinedFormat name ="dBase4" description ="dBase file format">
<part extension="dbf" mime="application/octet-stream"/>
<part extension="idx" mime="application/octet-stream"/>
</externallyDefinedFormat>
<externallyDefinedFormat name ="dBase4" description ="dBase file format">
<part extension="dbf" mime="application/octet-stream"/>
<part extension="idx" mime="application/octet-stream"/>
</externallyDefinedFormat>
<externallyDefinedFormat name="MSSQLServer7.0" description="MS SQL server version 7.0"/>
<archiveFormat name="zip" description="pkzip compressed archive format">
<part extension="zip" mime="application/zip"/>
</archiveFormat>

</dataFormats>
```

#3 - 12/17/2003 12:45 PM - Peter McCartney

ok the issue seems to be

1) we need a controlled enumeration for externallyDefinedFormat that is both recongized by users and parsable by applications

2) mime types were created to serve this purpose. Project alexandria investigated this and decided that a combination of both format name and mime types was needed, since the appropriate mime type is not always adquate. Read http://www.alexandria.ucsb.edu/middleware/dtds/ADL-access-report.dtd to see their discussion. Basically they provide three elements i their metadata schema for downloads - format, mime, and encoding.

3) vendors are slowly adding mime types but very few scientific data formats have been added. if we define mimes for these formats we could register them

only by putting an x- in front of it. and of course these definitions would be depracated when the owner puts in a definition.

4) dataFormat is required, so if the data are in Oracle, we need to have SOMETHING to put here, even if the information is superflous once connectionDefinition is filled out. its not clear to me if mimes even apply to connections - perhaps these are all octet-streams?

5) going beyond the enumeration issue, if we were to adopt a dictionary, we have the option of storing other metadata on a format that could be useful. the example i show here lists each part of a multipart format and its mime type. we use a file similar to this in our Xylopia data service to determine what parts of a file format need to be gathered up into the zip package. in my example it lists extensions which works fine for dealing with shapefiles, dbf, mapinfo, geoTiff, and so on. the only other multipart type that does not use extensions to identify its parts is arcinfo coverages. in this case the rules rely on foldernames and filenames under those folders to handle the different parts. because coverages within one folder share a common metadata folder, you can not move coverages by zipping up the files.. you must open it and save it as some other format for transport.

there was some debate about the utility of this multipart info, so im willing to table that part of the issue and continue to do it internally ourselves. but it would be really nice if we could agree how to ensure that shapefile will always be shapefile and not Shapefile, shape file, shape, esrishapefile...etc.

the attachment i put in (and edited) was an example of such a dictionary showing how multiprt, single part and service formats could all be handled using a strategy similar to ADA where we define format types, and then list the mimes for each of the parts. Matt felt this was inappropriate as there is in fact a multipart mime type.
so a variant on this would be to put the mime attribute in the externallyDefinedElement tag rather than in the part tag (or both). the nice ting about this is that like stmml.xsd, it abstracts users from complicated terminology yet does enable maching processing through mime types when they exist. if we leave the mime element out of eml, then the dictionary can have the most up-todate mime for any given format and we dont have to edit eml files when new mime types appear.

#4 - 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.

#5 - 03/27/2013 02:16 PM - Redmine Admin

Original Bugzilla ID was 1197

Files

formatDictionary2.xml	1.15 KB	12/17/2003	Peter McCartney