NORMALIZATION FOR XML 1.0 FILES:
A SPECIFICATION OF THE XML_NORM_1 FORMAT

Release Date: 06/05/2003

Author: Andrea Goethals, FCLA

Change History:

07/07/2003 Edited one entry (#6) and added two entries (#7, #8) to the list of locations within an XML document that references to external files can appear.

07/24/2003 1. Deleted all descriptions of the archive program's file resolution algorithm and instead refer to a separate FCLA publication for these details - “Resolving References in Distributed Objects”.
2. In the normalized versions file references are replaced with the file's DFID plus its file extension, instead of just its DFID.

XML_norm_1 format: A single XML file

Methodology (The underlined words in this section are defined in the *Glossary of Terms* section near the end of this paper):

Each XML 1.0 document that is submitted for archive ingest is parsed to see if it contains any file references to external files of any file format. If it does not, no normalized version is created for the XML document. If it contains at least one file reference, the archive program determines one-by-one if the references can be resolved. The algorithm used for file reference resolution is described in a separate FCLA report, “Resolving References in Distributed Objects” (available on the FCLA digital archive website). If at least one file reference can be resolved, a new XML document (the normalized version) is created.

In the new XML document, resolvable file references are substituted with the DFID plus the file extension of the referenced file. Any unresolvable file references to files are left 'as is' and these references are recorded in the archive database as 'broken links'. Broken links will not cause a package to be rejected from the ingest process.

There is the possibility of ambiguities in the file reference. An example of this occurring would be if the reference is in the form of a non-URL absolute path, i.e. C:\images\1.jpg. The absolute path to 1.jpg could have been valid within the file system of the computer that was used to create the XML file, but will not be valid for the archive program's file system. Ambiguities can also exist when the file reference is in the form of a relative path, i.e. images/1.jpg, when the file is not located in the expected place within the package. If the submitted XML 1.0 document is in METS format, the checksum values may be available for some file references, and can be used to resolve file reference ambiguities.

In many cases, the schema retrieved from the Internet will also be XML 1.0 documents. The archive program will parse them to see if they have any file references, but will use a different file resolution algorithm for these 'second tier' files as compared to the 'first tier' (submitted) XML documents. For details on these algorithms see “Resolving References in Distributed Objects”.

Detecting External References within XML Documents:

There are some syntactical forms of links to external files that will be detected for any XML documents by the archive program because they are commonly used in XML documents (see the Action Plan

Background of XML 1.0 Documents for examples of each):

1. A reference to an external DTD in a Document Type Declaration
2. The value of the schemaLocation attribute in the XMLSchema instance namespace
3. The value of the noNamespaceSchemaLocation attribute in the XMLSchema instance namespace
4. The value of the href attribute in the XLink namespace
5. A reference to an external file in a general entity declaration in an internal DTD
6. A reference to an external file in a parameter entity declaration in an internal DTD
7. A reference to an external file in a notation declaration in an internal DTD
8. The schemaLocation attribute of an import element in an XML Schema document
9. The schemaLocation attribute of an include element in an XML Schema document
10. The schemaLocation attribute of a redefine element in an XML Schema document
11. The value of the href attribute of a processing instruction whose target is xml-stylesheet
12. The value of the href attribute of an import element in an XSLT namespace
13. The value of the href attribute of an include element in an XSLT namespace
14. The parameter of the document function in an XSLT stylesheet
15. The value of the href attribute of an include element in an XInclude namespace

METS files use the above techniques for linking to external files, so all external links in METS files should be detected by the archive program (unless a schema 'extension' is used that doesn't conform to the above linking methods). MXF files add an MXF-specific technique for linking to external files: the filename element. The archive program will not be designed to detect MXF-specific linking techniques, because MXF files are not crucial to the archive's functions. If a user submits an MXF file as an initial package descriptor it is converted to a METS file prior to ingest into the archive.

Example 1: Submitted XML document that uses MXF schema:
The references to external files that would be detected by the ingest program for any XML document are underlined. The MXF-specific links that would not be recognized are in bold and italic font.

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>
<!DOCTYPE MXF SYSTEM "MXF.dtd">
<MXF>
    <package new="yes">
        <packageDesc id="UF00001644">
            <contrib creator="MXFClient, Sullivan, Mark">UF</contrib>
            <procInstr directory="D:\Antique\UF00001644\" makerules="map"
                   server="IC"></procInstr>
            <formats>image/jpeg, image/tiff 6.0, image/sid</formats>
            <timestamp>20021101T173117Z</timestamp>
        </packageDesc>
        <entityDesc id="UF00001644" type="map" source="UF">
            <projects>MAP, MAPAF</projects>
            <versionStatement>Electronic version created 2002, State
                   University Sytem of Florida.</versionStatement>
            <extRec pid="rk:AAA4208" sid="FCLANOTIS:QF"></extRec>
            <bibDesc ns="dc=http://purl.org/dc">
                <dc.title>AFRIQUE</dc.title>
            </bibDesc>
            <div type="main">
                <head>UF00001644</head>
                <div type="page">
                    <file format="image/jpeg" filesize="25455">
                        <filename>UF00001644.jpg</filename>
                        <checksum
                        type="MD5">a9dab149c191a402308d093a4d9ead49</checksum>
                        <creationDate>20021017T212942Z</creationDate>
                    </file>
                    <file format="image/sid" filesize="929354">
                        <filename>UF00001644.sid</filename>
                        <checksum
                        type="MD5">cf46f1218bef3f9ab8dc76c53904c71b</checksum>
                        <creationDate>20021017T201652Z</creationDate>
                    </file>
                    <file format="image/tiff 6.0" filesize="111031452">
                        <filename>UF00001644.tif</filename>
                        <checksum
                        type="MD5">ce2455862f9f327e2426e4d7c47abb73</checksum>
                        <creationDate>20021017T152638Z</creationDate>
                        <creationMethod software="Adobe Photoshop
                               7.0"></creationMethod>
                        <compression name="Uncompressed"></compression>
                        <image>
                            <bitDepth>8,8,8</bitDepth>
                            <storage segment="strip"
                               planarConfiguration="chunky"></storage>
                            <samplingFrequency unit="inch" x="450"
                               y="450"></samplingFrequency>
                            <colorSpace>RGB</colorSpace>
                            <dimensions x="7272" y="5088"></dimensions>
                        </image>
                    </file>
                </div>
            </div>
        </entityDesc>
    </package>
</MXF>
```

The above file has 1 external link that would be detected by the archive program:
1. MXF.dtd
If the archive program were able to locate the physical file associated with this name, a normalized version of this XML document would be created, replacing the link to MXF.dtd with a link to the DFID plus extension assigned to MXF.dtd.

Example 2: Submitted XML Document in METS format
The references to external files that would be detected by the ingest program for any XML document are underlined. Note that there are not any links to external files that would not be detected by the archive program in this example because the METS schema uses linking techniques that are common to XML documents.

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>
<METS:mets xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:METS="http://www.loc.gov/METS/"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xlink="http://www.w3.org/TR/xlink"
xmlns:palmm="http://www.fcla.edu/dls/md/palmm/"
xmlns:techmd="http://www.fcla.edu/dls/md/techmd/"
xmlns:rightsmd="http://www.fcla.edu/dls/md/rightsmd/"
xsi:schemaLocation="http://www.loc.gov/METS/
http://www.loc.gov/standards/mets/mets.xsd
http://purl.org/dc/elements/1.1/
http://dublincore.org/schemas/xmls/simpledc20021212.xsd
```

```
http://www.fcla.edu/dls/md/palmm/ http://www.fcla.edu/dls/md/palmm.xsd
http://www.fcla.edu/dls/md/techmd/ http://www.fcla.edu/dls/md/techmd.xsd
http://www.fcla.edu/dls/md/rightsmd/ http://www.fcla.edu/dls/md/rightsmd.xsd"
OBJID="UFE0000500"
TYPE="ETD"
LABEL="Data Mining Meets E-Commerce: Using Data Mining to Improve Customer
Relationship Management">

<METS:metsHdr CREATEDATE="2003-05-14T09:26:18Z" LASTMODDATE="2003-05-
14T09:26:18Z" ID="UFE0000500" RECORDSTATUS="NEW">
      <METS:agent ROLE="CREATOR" TYPE="ORGANIZATION">
            <METS:name>UFRGP</METS:name>
            <METS:note>projects=ETD</METS:note>
            <METS:note>makerules=title,UF</METS:note>
            <METS:note>server=TD</METS:note>
      </METS:agent>
</METS:metsHdr>
<METS:dmdSec ID="DMD1">
      <METS:mdWrap MIMETYPE="text/xml" MDTYPE="DC" LABEL="Simple Dublin Core">
            <METS:xmlData>
                  <dc:title>Data Mining Meets E-Commerce: Using Data Mining to
Improve Customer Relationship Management</dc:title>
                  <dc:date>2002</dc:date>
                  <dc:creator>Adderly, Darryl M.</dc:creator>
                  <dc:publisher>University of Florida</dc:publisher>
                  <dc:subject></dc:subject>
                  <dc:description>The application of data mining techniques to
the World Wide Web, referred to as Web mining, enables businesses to use
knowledge discovered from the past to understand the present and make critical
business decisions about the future. For example, this can be done by
analyzing the Web pages that visitors have clicked on, items that they have
selected or purchased, or registration information provided while browsing. To
perform this analysis effectively, businesses find the natural groupings of
users, pages, etc., by clustering data stored in the Web logs. The standard k-
means algorithm, an iterative refinement algorithm, is one of the most popular
clustering methods used today and it has proven to be an efficient clustering
technique. However, numerous iterations over the data set and re-calculating
cluster centroid values are time consuming. In this thesis, we improve the
time complexity of the standard algorithm. Our single-pass, non-iterative k-
means algorithm scans the data only once, calculating all the point and
centroid values based on the desired attributes of interest, and places the
items within their respective cluster thresholds. Our Web mining process
consists of three phases, pre-processing, pattern discovery, and pattern
analysis, which are described in detail in the thesis. We will use our
implementation of the k-means algorithm to uncover meaningful Web trends to
understand and, after analyzing the results, provide recommendations that may
have improved the visitor s website experience. We find that the clustering
results of our algorithm provide the same amount of knowledge for analysts as
one of the industry s leading data mining applications.  </dc:description>
                  <dc:language>EN</dc:language>
            </METS:xmlData>
      </METS:mdWrap>
</METS:dmdSec>

<METS:dmdSec ID="DMD2">
      <METS:mdWrap MIMETYPE="text/xml" MDTYPE="OTHER" LABEL="PALMM
Extensions">
      <METS:xmlData>
            <palmm:thesis>
                  <palmm:committeeChair>Hammer, Joachim
                   </palmm:committeeChair>
                  <palmm:committeeMember>Dankel, Douglas
                   D.</palmm:committeeMember>
                  <palmm:committeeMember>Lam, Herman </palmm:committeeMember>
                  <palmm:graduationDate>2003-12-21</palmm:graduationDate>
                  <palmm:degree>M.S.</palmm:degree>
                  <palmm:degreeDiscipline>Computer and Information Science and
                         Engineering</palmm:degreeDiscipline>
                  <palmm:degreeGrantor>University of
                         Florida</palmm:degreeGrantor>
                  <palmm:degreeLevel>Masters</palmm:degreeLevel>
            </palmm:thesis>
      </METS:xmlData>
</METS:mdWrap>
</METS:dmdSec>

<METS:amdSec>
      <METS:rightsMD ID="RMD1">
            <METS:mdWrap MDTYPE="OTHER" OTHERMDTYPE="RIGHTSMD">
                  <METS:xmlData>
                        <rightsmd:accessCode>public</rightsmd:accessCode>
                  </METS:xmlData>
            </METS:mdWrap>
      </METS:rightsMD>
      <METS:sourceMD ID="SMD1">
            <METS:mdWrap MIMETYPE="text/xml" MDTYPE="OTHER"
                   OTHERMDTYPE="PALMM">
                  <METS:xmlData>
                        <palmm:entityDesc SOURCE="UF"/>

                  </METS:xmlData>
            </METS:mdWrap>
      </METS:sourceMD>
</METS:amdSec>

<METS:fileSec>
      <METS:fileGrp ID="FG1">
            <METS:file ID="F1" MIMETYPE="application/pdf" SEQ="1"
                   CREATED="2003-04-16T11:00:47" SIZE="2072208"
                   CHECKSUM="682318CB81699762364D1E134B2AA2D8" ADMID="RMD1">
                  <METS:FLocat LOCTYPE="OTHER" OTHERLOCTYPE="SYSTEM"
                         xlink:href="adderly_d.pdf"/>
            </METS:file>
      </METS:fileGrp>
</METS:fileSec>
<METS:structMap ID="SM1">
      <METS:div TYPE="ETD" DMDID="DMD1">
            <METS:div ORDER="1" TYPE="main" LABEL="Data Mining Meets
            E-Commerce: Using Data Mining to Improve Customer Relationship
             Management">
                  <METS:fptr FILEID="F1"/>
            </METS:div>
      </METS:div>
</METS:structMap>

</METS:mets>
```

The above file has 6 links to external files:

1. http://www.loc.gov/standards/mets/mets.xsd
2. http://dublincore.org/schemas/xmls/simpledc20021212.xsd
3. http://www.fcla.edu/dls/md/palmm.xsd

4. http://www.fcla.edu/dls/md/techmd.xsd
5. http://www.fcla.edu/dls/md/rightsmd.xsd
6. adderly_d.pdf

Using the algorithms described in “Resolving References in Distributed Objects” the archive program would try to locate the physical files associated with the URIs listed above. If any of the files were found a new (normalized) version of the submitted XML document would be created replacing the URIs with the DFIDs plus the file extension for the located referenced files.

Glossary of Terms:

DFID = Unique identifier (a string of 4 upper-case alphanumeric characters followed by 4 unsigned numeric digits) assigned to each file by the archive program at the time when the file is ingested or created by the program.

File reference = A URI that is located in one of multiple particular locations within a document. For example, a relative path to an external DTD within an XML document's Document Type Declaration.

Package = A directory conforming to a particular naming scheme, located in a particular directory, containing all the submitted files related to a logical object.

Relative path = A file system path that uses another file or directory (a 'context') as a reference point in resolving the path. For example, the relative path “images/logo.jpg” found within the file “http://www.example.com/index.html” is resolvable to “http://www.example.com/images/logo.jpg”. Relative paths are distinguishable from 'absolute paths' which begin with a '/' on UNIX systems, and a drive (i.e. 'C:') on Windows systems. URLs are equivalent to absolute paths in that they do not need a context to be understood.

Resolved = the named reference can be associated with an actual physical file by the archive program.

URL = “Uniform Resource Locater” - the global address of WWW documents and resources. The first part of the address specifies the protocol (i.e. http or ftp), the second part species the domain name (i.e. www.fcla.edu) and the third part specifies the path to the resource (i.e. digitalArchive/images/logo.jpg). The possible protocols include http, ftp, gopher, mailto, news, telnet, wais, file, nntp, and prospero.
- The http URL has the form:
http://<host>:<port>/<path>?<searchpart>
The “:<port>” and “?<searchpart>” portions are optional.
- The ftp URL has the form:
ftp://<user>:<password>@<host>:<port>/<path>
The “<user>:<password>@”, “:<password>”, “:<port>”, and “/<path>” portions are optional.

XML Schema language instance = A document that can be used to determine if a class of XML documents are valid. Schema languages include DTDs, W3C XML Schema, RELAX NG, Schematron, and Examplotron.

References:

Baker, David W. “A Guide to URLs”, 1996. http://www.netspace.org/users/dwb/url-guide.html

Berners-Lee, T.; L. Masinter, and M. McCahill. “Uniform Resource Locators (URLs)”, RFC: 1738, December 1994. http://www.cis.ohio-state.edu/cgi-bin/rfc/rfc1738.html

Goethals, Andrea. “Action Plan Background: XML 1.0”, FCLA Digital Archive website, June 9, 2003.

Van der Vlist, Eric. “Comparing XML Schema Languages”, O'Reilly XML.com; O'Reilly & Associates, Inc.; December 12, 2001. http://www.xml.com/pub/a/2001/12/12/schemacompare.html

Vickery, Chris and Andrea Goethals. “Resolving References in Distributed Objects”, FCLA Digital Archive website, July 23, 2003.