18. XML BIBLIOGRAPHIC HEADER DESCRIPTION

The purpose of this guideline is to describe the Extensible Markup Language (XML) format that will be used to structure the LSN document bibliographic header data in a canonical form. It also provides guidance on how the data is stored in a file. The Design Requirements describe the use and values of the fields. Use of a common format will facilitate the exchange of data between the central LSN website and LSN participants.

This information is being presented in a way that can be understood by those not familiar with XML. For an additional overview and introduction to XML, please follow the link below:

http://www.ed.gov/databases/ERIC_Digests/ed437941.html

18.1 Header Generator Tool

Participants who do not have many documents to load on their LSN server, and who do not already have bibliographic headers created, may wish to use the header generator tool provided on the LSN. After logging in, click on Help on the global navigation bar at the top. From the Help home page, click on “Other Downloads” under the Other Help section of the local navigation bar on the left. The header generator tool is located at the top, under LSN Support Software. Click on the link for the appropriate operating system to download the self-extracting file for installation. E-mail LSNWebmaster@nrc.gov with any questions.

18.2 XML Structure Overview

The basic structure of XML is as follows:

```
<?xml version="1.0" ?>
        <Headers>
                <Header>
                .....
                </Header>
                <Header>
                .....
                </Header>
        </Headers>
```

One file can contain more than one bibliographic header record; however, even though the LSN software can handle more than one record per file, it is suggested that LSN participants use one record per file.

The LSN format uses mixed-case tags, e.g., <DocumentNum>. Furthermore, the XML tags are case-sensitive. Therefore, <DOCUMENTNUM> or <documentnum> are not the same as <DocumentNum>.

Some of the fields in the bibliographic header can contain multiple values. Tags that could contain multiple entries are plural. For example, Author**s** indicates that any given document has the potential for more than one author. When a document has only one author, the tagging scheme remains the same. For example, if a document has two authors it is tagged as follows:

```
<Authors>
        <Author Org="SNL">Smith JR</Author>
        <Author Org="SNL">George JT</Author>
</Authors>
```

If a document has only one author, it is still wrapped on the <Authors> tag:

```
<Authors>
        <Author Org="SNL">Smith JR</Author>
</Authors>
```

Refer to the Design Requirements for a listing of which attributes are multi-valued.

18.3 Special Characters

Five characters cannot be present in XML data without special handling. They are: & < > ’ ". There are two ways to handle data containing any of these special characters. The first method is to “escape” the character. The second method is to mark up sections with CDATA (character data).

For example, assume the title of the document is “Jack & Jill.” If the ampersand in the title is left as is, XML will generate an error:

```
<Title>Jack & Jill</Title>
```

As noted above, one method to easily correct this is to “escape” the character. Using escape to correctly handle the ampersand, the correct title is:

```
<Title>Jack &amp; Jill</Title>
```

When the results of a search in the LSN are displayed, the title will appear correctly as “Jack & Jill.”

The escape codes for the five characters are provided in the following table:

Character	Escape Code
&	&amp;
<	&lt;
>	&gt;
’	&apos;
"	&quot;

The second method of handling data containing the five special characters is to use the CDATA sections. The term CDATA is inherited from SGML (Standard Generalized Markup Language). CDATA encapsulates all characters in a value and tells the XML parser to ignore them. When the CDATA section is used, the five special characters do not need to be escaped. An example using CDATA is:

<Title><**![CDATA[Jack & Jill]]**</Title>

The CDATA section can only be used for XML elements. An XML element is text between a start tag (e.g., <TITLE>) and an end tag (e.g., </TITLE>). The CDATA section cannot be used to encode special characters in element attributes (the text within an attribute). For format purposes, an element attribute is an assignment statement within the start tag of an element. An example of an escape code being applied to the Author Organization “AT&T” is:

<Author Org=”AT&amp;T”/>

The discussion above detailed how the special character ampersand is handled. The same methodology is followed for the remaining special characters < > ‘ “. When a special character occurs in the text between a tag, it must be escaped with an escape sequence or with a CDATA section. When these characters occur in an attribute value, they must be escaped with an escape sequence.

18.4 Sample XML Bibliographic Header Layout

The following is a sample XML bibliographic header layout for LSN participants. The data is provided as a technical example and does not represent an actual document. The data represented in the XML bibliographic headers is analogous to the following catalog format example:

Excess Control Information:
Addressee Name:
Addressee Organization:
Author Name:
Author Organization:

Comments:
Descriptors:
Document Date:
Document Number(s):
Document Type:
Image URL:
Non-Digital Media Indicator:
Number of Images:
Package Identifier:
Participant Accession Number:
A Record Indicator:
Related Record Code:
Related Record Number:
Text URL:
Title [Created Title]:
Traceability Code and Number:
Version:

The example below is the format for an LSN XML encoded a bibliographic header. All of the bibliographic header fields defined in the Design Requirements are presented. However, not all attributes are required; some are optional. Please refer to the Design Requirements for a listing of which attributes are optional. If a tag is optional, omitting it will make the bibliographic header file smaller. Note also that some elements contain multiple values (i.e., Authors). The indentations convey structure and some lines wrap due to their length. When a line appears to be improperly indented, it is because of wrapping from the line above it.

URL paths do not need to be fully qualified; they can be relative. The application will navigate relative paths to find subdirectories.

```
<?xml version="1.0" ?>
<Headers>
    <Header>
        <ParticipantAccNum>MOL.19961111.0013</ParticipantAccNum>
        <DocDate>19960603</DocDate>
        <NonDigitalMedia> </NonDigitalMedia>
        <NumOfImages>1</NumOfImages>
        <QARecInd>Y</QARecInd>
        <Title>Change FINAL PUBLISHED DOCUMENT FOR SAND94-2322, BENCH-SCALE
EXPERIMENTAL DETERMINATION OF THE THERMAL DIFFUSIVITY OF CRUSHED TUFF,
JUNE 1, 1996 (C)</Title>
        <Descriptors>EXPERIMENTAL TURF</Descriptors>
        <Comment>This document is a sample of the DOE1.XML file to be used only for LSN
testing/development. The DOE does not certify the content, adequacy or technical accuracy of
the material.</Comment>
        <URLs>
            <URL Type="T" Page="0">http://nrc/MOL199611110011EX1.HTM</URL>
            <URL Type="I" Page="1">http:/image/MOL199611110011P00001.TIF</URL>
        </URLs>
        <Authors>
            <Author Org="SNL"><![CDATA[O'Reilly RE]]></Author>
            <Author Org="SNL">George JT</Author>
        </Authors>
        <Addressees>
            <Addressee Org="NRC">Turner JX</Addressee>
        </Addressees>
        <Traceabilities>
            <Traceability>1.2.1.5</Traceability>
        </Traceabilities>
        <AccCtrls>
            <AccCtrl>PRV</AccCtrl>
        </AccCtrls>
        <DocumentNums>
            <DocumentNum>SAND94-2322</DocumentNum>
        </DocumentNums>
        <DocumentTypes>
            <DocumentType>Publication</DocumentType>
        </DocumentTypes>
        <RelatedRecs>
            <RelatedRec Code="AMR">MOL.20000113.0488</RelatedRec>
        </RelatedRecs>
        <Versions>
            <Version>1</Version>
        </Versions>
        <PackageIds>
            <PackageId>MOY-961126-34-02</PackageId>
        </PackageIds>
    </Header>
</Headers>
```

18.5 LSN Accession Number Assignment

Participants do not have to assign the master accession number that will be used by the LSN software. The accession number is assigned by the LSN software the first time an object file is encountered by the "crawler." The LSN accession number is retained only on the central LSN portal server and it does ***not*** have to be stored locally, appended to, or otherwise linked to by the participants on their servers. The central LSN website will maintain the originally assigned LSN accession number. Even if an item subsequently is removed from a participant LSN site, the LSNA will maintain a "transactions" log on the central LSN website so that participants can review and account for all used LSN accession numbers.