DDI TAG DOCUMENT

Section 3.0 – DATA FILES DESCRIPTION

Tags in this section relate directly to the format and content of the datafiles. Note that the tags should correspond to the format you are working with – raw datafile, SAS dataset, NSDstat file.

"The File Description consists of information about the particular data file(s) containing numeric and/or numeric + textual information that the DDI-compliant file describes. This section consists of items describing the characteristics and contents of file(s) that comprise the study as described in the Study Description. There may be multiple file descriptions if there are multiple files in the collection."

Source: DDI Codebook

DTD Numbers	Tags
3.0	<fileDscr>
3.1	<fileTxt>
3.1.1	<fileName>
3.1.2	<fileCont>
3.1.3	<fileStrc>
3.1.3.1	<recGrp >
3.1.4	<dimensns>
3.1.4.1	<caseQnty>
3.1.4.2	<varQnty>
3.1.4.3	<logRecL>
3.1.4.5	<recNumTot>
3.1.5	<fileType>
3.1.6	<format>
3.1.8	<dataChck>
3.1.12	<verStmt>
3.1.12.1	<version>
3.1.12.2	<verResp>
3.1.12.3	<notes >
3.3	<notes >

3.0 <fileDscr> Data File Description

- Optional
- Repeatable
- Attributes: ID, xml:lang, source, URI, sdatrefs, methrefs, pubrefs, access

Description: Information about the data file(s) that comprises a collection. This section can be repeated for collections with multiple files. The "URI" attribute may be a URN or a URL that can be used to retrieve the file. The "sdatrefs" are summary data description references that record the ID values of all elements within the summary data description section of the Study Description that might apply to the file. These elements include: time period covered, date of collection, nation or country, geographic coverage, geographic unit, unit of analysis, universe, and kind of data. The "methrefs" are methodology and processing references that record the ID values of all elements within the study methodology and processing section of the Study Description that might apply to the file. These elements include information on data collection and data appraisal (e.g., sampling, sources, weighting, data cleaning, response rates, and sampling error estimates). The "pubrefs" attribute provides a link to publication/citation references and records the ID values of all citations elements within Other Study Description Materials or Other Study-Related Materials that pertain to this file. "Access" records the ID values of all elements in the Data Access section that describe access conditions for this file.

Remarks: When a codebook documents two different physical instantiations of a data file, e.g., logical record length (or OSIRIS) and card-image version, the Data File Description should be repeated to describe the two separate files. An ID should be assigned to each file so that in the Variable section the location of each variable on the two files can be distinguished using the unique file IDs

Example(s):

<fileDscr ID="**CARD-IMAGE" URI="www.icpsr.umich.edu/cgi-bin/archive.prl?path=ICPSR&num=7728"**/>

<fileDscr ID**="LRECL" URI="www.icpsr.umich.edu/cgi-bin/archive.prl?path=ICPSR&num=7728"**/>

3.1 <fileTxt> File Description

- Optional
- Repeatable
- Attributes: ID, xml:lang, source

Description: Information about the data file. Demonstrates that the next section of the codebook will deal with the datafile.

3.1.1 <fileName> File Name

- Optional
- Not Repeatable
- Attributes: ID, xml:lang, source

Description: Contains a short title that will be used to distinguish a particular file/part from other files/parts in the data collection.

<fileName>**Youth Smoking Survey, 2002**</fileName>

3.1.2.1 <fileCont> Contents of Files

- Optional
- Not Repeatable
- Attributes: ID, xml:lang, source

Description: Abstract or description of the file. A summary describing the purpose, nature, and scope of the data file, special characteristics of its contents, major subject areas covered, and what questions the PIs attempted to answer when they created the file. A listing of major variables in the file is important here. In the case of multi-file collections, this uniquely describes the contents of each file.

Example:

<fileCont>**Part 1 contains both edited and constructed variables describing demographic and family relationships, income, disability, employment, health insurance status, and utilization data for all of 1987.**</fileCont>

<fileCont>**The annual HIUS contains detailed data on the Internet activities of Canadian household.** </fileCont>

3.1.3 <fileStrc> File Structure

- Optional
- Not Repeatable
- Attributes: ID, xml:lang, source, type

Description: Type of file structure. Use attribute of "type" to indicate hierarchical, rectangular, or relational (the default is rectangular).

Example:

<fileStrc>**Rectangular**</fileStrc>

<fileStrc>**Hierarchical**</fileStrc>

3.1.3.1 <recGrp > Record or Record Group

- Optional
- Repeatable
- Attributes: ID, xml:lang, source, recGrp, rectype, keyvar, rtypeloc, rtypewidth, rtypevtype, recidvar

Description: Used to describe record groupings if the file is hierarchical or relational. The attribute "recGrp" allows a record group to indicate subsidiary record groups that nest underneath; this allows for the encoding of a hierarchical structure of record groups. The attribute "rectype" indicates the type of record

Example:

<recGrp>**Person-level records**</recGrp>

3.1.4 <dimensns> File Dimensions

- Optional
- Not Repeatable
- Attributes: ID, xml:lang, source

Description: Dimensions of the overall file.

3.1.4.1 <caseQnty> Number of cases / Record Quantity

- Optional
- Repeatable
- Attributes: ID, xml:lang, source

Description: Number of cases or observations in the entire file. This is to be used for rectangular files only.

Example: (from SHS 2001)

<caseQnty>**16901**</caseQnty>

3.1.4.2 <varQnty> Number of variables per record

- Optional
- Repeatable
- Attributes: ID, xml:lang, source

Description: Number of variables in the entire file. This is to be used for rectangular files only.

Example: (from SHS 2001)

<varQnty>**255**</varQnty>

3.1.4.3 <logRecL> Record Length / Logical Record Length

- Optional
- Repeatable
- Attributes: ID, xml:lang, source

Description: Logical record length of the file, i.e., number of characters of data in the record. Only to be used for rectangular files or if all records in a hierarchical file are the same length.

Example: (from SHS 2001 raw data file)

<logRecL>**2093**</logRecL>

3.1.4.5 <recNumTot> Overall Number of Records

- Optional
- Repeatable
- Attributes: ID, xml:lang, source

Description: Overall record count in the file. Particularly helpful in instances such as files with multiple cards/decks or records per case.

Example:

<recNumTot>**2400**</recNumTot>

3.1.5 <fileType> Type of File

- Optional
- Not Repeatable
- Attributes: ID, xml:lang, source, charset

Description: Types of data files include raw data (ASCII, EBCDIC, etc.) and software-dependent files such as SAS datasets, SPSS export files, etc. If the data are of mixed types (e.g., ASCII and packed decimal), state that here. Note that the element varFormat permits specification of the data format at the variable level. The "charset" attribute allows one to specify the character set used in the file, e.g., US-ASCII, EBCDIC, UNICODE UTF-8, etc.

Example:

<fileType>**shs2001.NSDstat**</fileType> for a Nesstar datafile
<fileType>**shs2001.sas7dbat**</fileType> for a SAS dataset
<fileType>**PUMDFSHS2001.txt**</fileType> for a raw datafile

3.1.6 <format> Data Format

- Optional
- Not Repeatable
- Attributes: ID, xml:lang, source

Description: Physical format of the data file: Logical record length format, card-image format, de-limited format, free format, etc.

Example:

<format>**comma-delimited**</format>

3.1.8 <dataChck> Extent of Processing Checks

- Optional
- Not Repeatable
- Attributes: ID, xml:lang, source

Description: Indicate here at the file level the types of checks and operations performed on the data file. A controlled vocabulary may be developed for this element in the future.

Example:

<dataChck>**Consistency checks were performed by the Data Archives.**</dataChck>

3.1.12 <verStmt> Version Statement

- Optional
- Not Repeatable
- Attributes: ID, xml:lang, source

Description: Version statement for the data file, if one of a multi-file collection.

3.1.12.1<version> Version

- Optional
- Not Repeatable
- Attributes: ID, xml:lang, source, date, type

Description: Also known as release or edition. If there have been substantive changes in the file since its creation, this statement should be used. The ISO standard for dates (YYYY-MM-DD) is recommended for use with the date attribute.

Example:
<version type='**revision**' date=**'2004-02-05'**>**Second Revision of SHS data**</version>

3.1.12.2 <verResp> Version Responsibility Statement

- Optional
- Not Repeatable
- Attributes: ID, xml:lang, source, affiliation

Description: Used to indicate the organization or person responsible for the version of the file.

Example:

<verResp> **Statistics Canada, Income Statistics Division**</verResp>

3.1.12.3 <notes> Notes

- Optional
- Repeatable
- Attributes: ID, xml:lang, source, type, subject, level, resp

Description: Used to indicate additional information regarding the version or version responsibility statement, in particular to indicate what makes a new version different from its predecessor. The attributes for notes permit a controlled vocabulary to be developed ("type" and "subject"), indicate the "level" of the DDI to which the note applies (study, file, variable, etc.), and identify the author of the note ("resp").

<notes>**Refer to shs1997-2000rdi.pdf for comparison of previously published data, Canada, 1997 to 2000.**</notes>

3.2 <notes> Notes

- Optional
- Repeatable

- Attributes: ID, xml:lang, source, type, subject, level, responsibility

Description: Additional information about the data file not covered in other elements. "Notes" sections appear in several places in the DDI. The attributes for notes permit controlled vocabulary to be developed (type and subject), the level of the DDI to which the note refers to be identified (study, file, variable, etc.), and the author of the note to be indicated (resp).

<notes>**There is a restricted version of this file containing confidential information, access to which is controlled by the principal investigator**.</notes>

Example 1:

SES 1996

```
- <fileDscr>
  - <fileTxt>
      <fileName>Sun Exposure Survey 1996</fileName>
      <fileCont>The data file contains information regarding the attitudes and behaviours related to the sun exposure amon
        over.</fileCont>
    - <fileStrc type="Rectangular">
      - <recGrp>
        - <recDimnsn>
            <varQnty />
            <caseQnty />
            <logRecL />
          </recDimnsn>
        </recGrp>
      </fileStrc>
    - <dimensns>
        <caseQnty>4,023</caseQnty>
        <varQnty>151</varQnty>
        <logRecL>188</logRecL>
        <recNumTot>1208</recNumTot>
      </dimensns>
    - <verStmt>
        <version type="revision" date="2001-06-15">2nd Revision of SES data</version>
        <verResp>Special Surveys Division, Statistics Canada</verResp>
        <notes>Users should be aware that Statistics Canada updated the datafiles for this study in July, 2001</notes>
      </verStmt>
    </fileTxt>
  </fileDscr>
```