DDI TAG DOCUMENT

Section 4.0 – VARIABLES (DATA) DESCRIPTION

Tags in this section relate to the variables and any statistic calculated from the data.

“The Variable Description consists of a section describing variable groups and a section describing individual variables within the data file. The variable groups are defined as variables that may share common subjects, arise from the interpretation of a single question, or are linked by some other factor. The Variable Description is an extremely rich set of elements allowing for detailed descriptive information about response and analysis units, question text, forward progression and back flow, interviewer instructions, universe, valid and invalid data ranges, derived variables, summary statistics, etc. References to other parts of the DDI-compliant documentation file are possible through the use of IDREFS and links so that interrelationships among the elements may be used and documented.”

Source: DDI Codebook

DTD Numbers	Tags
4.0	<dataDscr>
4.1	<varGrp>
4.1.1	<labl >
4.2	<var>
4.2.1	<location>
4.2.2	<labl>
4.2.8	<qstn>
4.2.8.1	<preQTxt>
4.2.8.2	<qstnLit>
4.2.8.3	<postQTxt>
4.2.8.6	<ivuInstr>
4.2.9	<valrng>
4.2.10	<invalrng>
4.2.12	<universe>
4.2.18	<catgry>
4.2.18.1	<catValu>
4.2.18.2	<labl>
4.2.18.4	<catStat>

Description of tags and working examples

4.0 <dataDscr> Variable Description

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

Description: Shows that the section dealing with the data is beginning.

4.1 <varGrp> Variable Group

- Optional
- Repeatable
- Attributes: ID, xml:lang, source, type, var, varGrp, name, sdatrefs, methrefs, pubrefs, access, nCube

Description: A group of variables that may share a common subject, arise from the interpretation of a single question, or are linked by some other factor.

Attributes within the <varGrp>

ID
To uniquely identify the Variable Group

Type
The "type" of group attribute refers to the general type of grouping of the variables, e.g., subject, multiple response.

Specific variable groups, included within the 'type' attribute, are:
Section: Questions which derive from the same section of the questionnaire, e.g., all variables located in Section C.
Multiple response: Questions where the respondent has the opportunity to select more than one answer from a variety of choices, e.g., what newspapers have you read in the past month (with the respondent able to select up to five choices).
Grid: Sub-questions of an introductory or main question but which do not constitute a multiple response group, e.g., I am going to read you some events in the news lately and you tell me for each one whether you are very interested in the event, fairly interested in the fact, or not interested in the event.
Display: Questions which appear on the same interview screen (CAI) together or are presented to the interviewer or respondent as a group.
Repetition: The same variable (or group of variables) which are repeated for different groups of respondents or for the same respondent at a different time.
Subject: Questions which address a common topic or subject, e.g., income, poverty, children.
Version: Variables, often appearing in pairs, which represent different aspects of the same question, e.g., pairs of variables (or groups) which are adjusted/unadjusted for inflation or season or whatever, pairs of variables with/without missing data imputed, and versions of the same basic question.
Iteration: Questions that appear in different sections of the data file measuring a common subject in different ways, e.g., a set of variables which report the progression of respondent income over the life course.
Analysis: Variables combined into the same index, e.g., the components of a calculation, such as the numerator and the denominator of an economic statistic.
Pragmatic: A variable group without shared properties.
Record: Variable from a single record in a hierarchical file.
File: Variable from a single file in a multifile study.

Randomized: Variables generated by CAI surveys produced by one or more random number variables together with a response variable, e.g, random variable X which could equal 1 or 2 (at random) which in turn would control whether Q.23 is worded "men" or "women", e.g., would you favor helping [men/women] laid off from a factory obtain training for a new job?
Other: Variables which do not fit easily into any of the categories listed above, e.g., a group of variables whose documentation is in another language.

The "varGrp" attribute is used to reference all the subsidiary variable groups which nest underneath the current varGrp. This allows for encoding of a hierarchical structure of variable groups.

The attribute "name" provides a name, or short label, for the group.

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 group. 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 which record the ID values of all elements within the study methodology and processing section of the Study Description which might apply to the group. 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 Section 2.5 or Section 5.0 that pertain to this variable group.

The "access" attribute records the ID values of all elements in Section 2.4 of the document that describe access conditions for this variable group.

Var
Listing of variables that make up the Variable Group

Example:

<varGrp ID="**VG1F1**" type="**subject**" var="**V1 V2 V3 V4 V5 V6**">

Variables V1 through to V6 all deal with a common topic or subject area

4.1.1 <labl > Label

- Optional
- Repeatable
- Attributes: ID, xml:lang, source, level, vendor, country, sdatrefs

Description: A short description of the variable group.

The following labels are suggested when grouping variables where possible:

- Demographics (includes: age, sex, martial status, household composition, religion, language and mobility)
- Place of work, commuting
- Main activity, occupation
- Income

- Education
- Aboriginal
- Immigration
- Ethnicity, visible minority
- Dwelling
- Health and activity limitations
- Labour force activity
- Derived variables

Derived variables should be included under the groups to which they refer and under a separate grouping for derived variables only. If a variable does not easily fit into a variable grouping, check the front of the codebook for one. Additional groupings can be added to this list for exceptional variables through consultation with the CANDDI group.

In the variable label, the length of this phrase may depend on the statistical analysis system used (e.g., some versions of SAS permit 40-character labels, while some versions of SPSS permit 120 characters), although the DDI itself imposes no restrictions on the number of characters allowed. A "level" attribute is included to permit coding of the level to which the label applies, i.e. record group, variable group, variable, category group, category, nCube group, nCube, or other study-related materials. The "vendor" attribute was provided to allow for specification of different labels for use with different vendors' software. The attribute "country" allows for the denotation of country-specific labels. The "sdatrefs" attribute records the ID values of all elements within the Summary Data Description section of the Study Description that might apply to the label. These elements include: time period covered, date of collection, nation or country, geographic coverage, geographic unit, unit of analysis, universe, and kind of data.

Example:

<labl>**Location**</labl>

The variables above V1-V6 have been grouped and have been labeled as Location. These should match up with Codebook information provided by the Authoring Division.

4.2 <var> Variable

- Optional
- Repeatable
- Attributes: ID, xml:lang, source, name, wgt, wgt-var, weight, qstn, files, vendor, dcml, intrvl, rectype, sdatrefs, methrefs, pubrefs, access, aggrMeth, measUnit, scale, origin, nature, additivity, temporal, geog, geoVocab, catQnty

Description: This element describes all of the features of a single variable in a social science data file. There are a number of attributes within this tag that allows us to specify items such as the name of the variable, what the weight variable is, etc….

Attributes within the <var>

ID
To uniquely identify the Variable

Name
Name of the variable

Files
Indicates the file where the variable can be found

Wgt

Indicates whether the variable is the weight variable or not

Dcml
Indicates the number of decimal places that are held in the variable

Intrvl
Indicates whether the variable is continuous or discrete

Example:
<var ID="**V1**" name="**CASEID**" files="**F1**" dcml="**0**" intrvl="**contin**">
Variable V1 has a name of Caseid, has 0 decimals and is a continuous variable

<var ID="**V2**" name="**WEIGHT**" wgt="**wgt**" files="**F1**" dcml="**0**" intrvl="**discrete**">
Variable V2 has the name WEIGHT and has been designated as the Weight variable.

4.2.1 <location> Variable Location

- Optional
- Repeatable
- Attributes: ID, xml:lang, source, StartPos, EndPos, width, RecSegNo, fileid, locMap

Description: This is an empty element containing only the attributes listed below. Attributes include "StartPos" (starting position of variable), "EndPos" (ending position of variable), "width" (number of columns the variable occupies), "RecSegNo" (the record segment number, deck or card number the variable is located on), and "fileid" (an IDREF link to the fileDscr element for the file that this location is within). The fileid is necessary in cases where the same variable may be coded in two different files, e.g., a logical record length type file and a card image type file. Note that if there is no width or ending position, then the starting position should be the ordinal position in the file, and the file would be described as free-format.

Attributes within the <location>

StartPos
Indicates the starting position for the variable in the Datafile

EndPos
Indicates the ending position for the variables in the Datafile

Width
Indicates the width of the variable in the Datafile.

Example:

<location StartPos="**9**" EndPos="**9**" width="**1**" />
The variable starts in column 9 and ends in column 9 with a width of 1

<location StartPos="**10**" EndPos="**17**" width="**8**" />
The variable starts in column 10 and ends in column 17 with a width of 8

4.2.8 <qstn> Question

- Optional
- Repeatable
- Attributes: ID, xml:lang, source, qstn, var, seqNo, sdatrefs

Description: The question element may have mixed content. The element itself may contain text for the question, with the subelements being used to provide further information about the question. Alternatively, the question element may be empty and only the subelements used. The element has a unique question ID attribute which can be used to link a variable with other variables where the same question has been asked. This would allow searching for all variables that share the same question ID perhaps because the questions was asked several times in a panel design.

4.2.8.1 <preQTxt> PreQuestion Text

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

Description: Text describing a set of conditions under which a question might be asked.

Example:

<preQTxt> **This survey deals with various aspects of your health. I'll be asking about such things as physical activity, social relationships and health status. By health, we mean not only the absence of disease or injury but also physical, mental and social well-being**. </preQTxt>

4.2.8.2 <qstnLit> Literal Question

- Mandatory
- Not Repeatable
- Attributes: ID, xml:lang, source, sdatrefs

Description: Text of the actual, literal question asked.

Example:
<qstnLit> **During June to August, when you were at work, how much time each day (on average) were you in the sun?** </qstnLit>

4.2.8.3 <postQTxt> PostQuestion Text

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

Description: Text describing what occurs after the literal question has been asked.

Example:

<postQTxt>**Go to next module**. </postQTxt>

4.2.8.6 <ivuInstr> Interviewer Instructions

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

Description: Specific instructions to the individual conducting an interview.

Example:

<ivuInstr> **Read list. Mark all that apply.** <**/**ivuInstr>

4.2.9 <valrng> Range of Valid Data Values

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

Description: This is the actual range. The "UNITS" attribute of Range permits the specification of integer/real numbers. The"min" and "max" attributes specify values which are considered part of the range.

Example:

<valrng><range min="1" max="3" /> </valrng>

4.2.10 <invalrng> Range of Invalid Data Values

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

Description: Values for a particular variable that represent missing data, not applicable responses, etc.

Example:

<invalrng><range UNITS="INT" min="98" max=98" max="99"> </range>
<key>
98 Don't Know
99 Not stated
</key> </invalrng>

4.2.12 <universe>

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

Description: The group of persons or other elements that are the object of research and to which any analytic results refer. Age, nationality, and residence commonly help to delineate a given universe, but any of a number of factors may be involved, such as sex, race, income, veteran status, criminal convictions, etc. The universe may consist of elements other than persons, such as housing units, court cases, deaths, countries, etc. In general, it should be possible to tell from the description of the universe whether a given individual or element (hypothetical or real) is a member of the population under study. A "level" attribute is included to permit coding of the level to which universe applies, i.e., the study level, the file level (if different from study), the record group, the variable group, the nCube group, the variable, or the nCube level. The "clusion" attribute provides for specification of groups included (I) in or excluded (E) from the universe. If all the variables/nCubes described in the data documentation relate to the same population, e.g., the same set of survey respondents, this element would be unnecessary at data description level. In this case, universe can be fully described at the study level.

Example:

<universe clusion="**I**">**Individuals 15-19 years of age.**</universe>

<universe clusion="**E**">**Individuals younger than 15 and older than 19 years of age.**</universe>

4.2.18 <catgry> Category

- Optional
- Repeatable
- Attributes: ID, xml:lang, source, missing, mistype, country, sdatrefs,excls

Description: A description of a particular response. The attribute “missing” indicates whether this category group contains missing data or not. The attribute “missType” is used to specify the type of missing data, e.g. inap., don’t know, no answer, etc. The attribute “country” allows for the denotation of country-specific category values. The “sdatrefs” attribute records the ID values of all elements within the summary data description that apply to this category. The exclusiveness attribute (“excls”) should be set to “false” if the category can appear in more than one place in the classification hierarchy.

4.2.18.1 <catValu> Category Value

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

Description: The explicit response.

Example:
<catValu>**24**</catValu>

The value for a given category in Var X is 24.

OR
<var><catgry missing=”Y” missType=”inap”><catValu>9</catValu></catgry></var>

4.2.18.2 <labl> Label

- Optional
- Repeatable
- Attributes: ID, xml:lang, source, level, vendor, country, sdatrefs

Description: A short description of the response. In the variable label, the length of this phrase may depend on the statistical analysis system used (e.g., some version of SAS permit 40-character labels, while some versions of SPSS permit 120 characters). Although the DDI itself imposes no restrictions on the number of characters allowed. A “level” attribute is included to permit coding of the level to which the label applies, i.e. record group, variable group, variable, category group, category, nCube group, nCube, or other study-related materials. The “vendor” attribute was provided to allow for specification of different labels for use witj different vendors’ software. The attribute “country” allows for the denotation of country-specific labels. The “sdatrefs” attribute records the ID values of all elements within the Summary Data Description section of the Study Description that might apply to the label. These elements include: time period covered, date of collection, nation or country, geographic coverage, geographic unit, unit of anaysis, universe, and kind of data.

Example:
<labl>**Always**</labl>

<labl>**Not in the labour force**</labl>

<labl>**Marital status of head of household**</labl>

The label for the category in Var X is “Always”

4.2.18.4 <catStat> Category Group Statistics

- Optional
- Repeatable
- Attributes: ID, xml:lang, source, type, URI, methrefs, wgtd, wgt-var, weight, sdatrefs

Description: May include frequencies, percentages, or cross-tabulation results which define the category; often appears in a table. This field can contain one of the following: 1. textual information (e.g. PCDATA), or 2. non-parseable character data (e.g. the statistics), or 3. some other form of external information (table, image, etc.). In case 1, the tag can be used to mark up character data; tables can also be included in the actual markup. In cases 2 or 3, the element can be left empty and the “URI” attribute used to refer to the external object containing the information. The attribute “type” indicates the type of statistics presented – frequency, percent, or crosstabulation.

Attributes within the <catStat>

Type
The attribute "type" refers to "frequency", "percent", or "crosstab".

Example:
<catStat type="**freq**">**16385**</catStat>

A frequency was calculated with a value of 16385

4.2.24 <notes> Notes

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

Description: Used to indicate additional information regarding the variable. “Notes” sections appear in several places in the DTD. The attributes for notes permit a controlled vocabulary to be developed (type and subject), the level of the DTD to which the note refers to be identified (study, file, variable, etc.), and the author of the note to be indicated (resp).

Example:

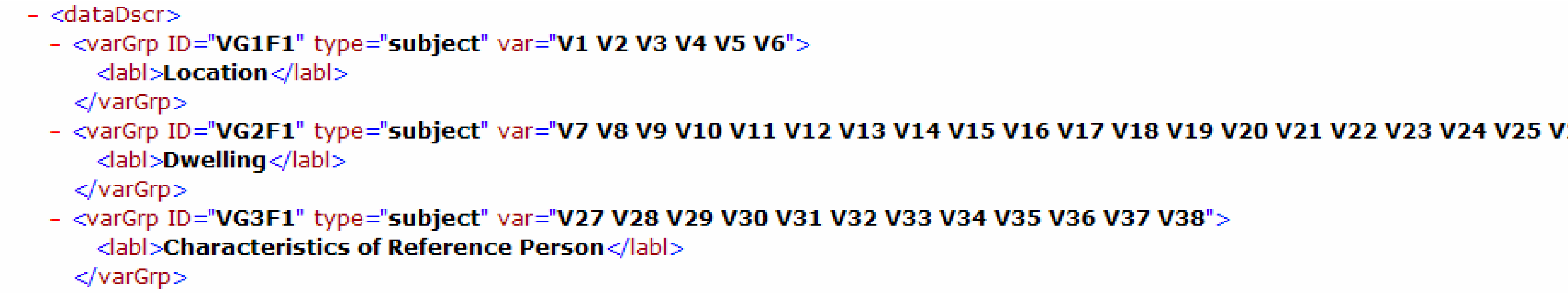
<notes>**Derived roster ages of household members. On public use microdata file, the maximum number of the members has been set at 02**.</notes>

<notes>**On public use microdata file, ages greater than 85 were recoded to 85**.</notes>

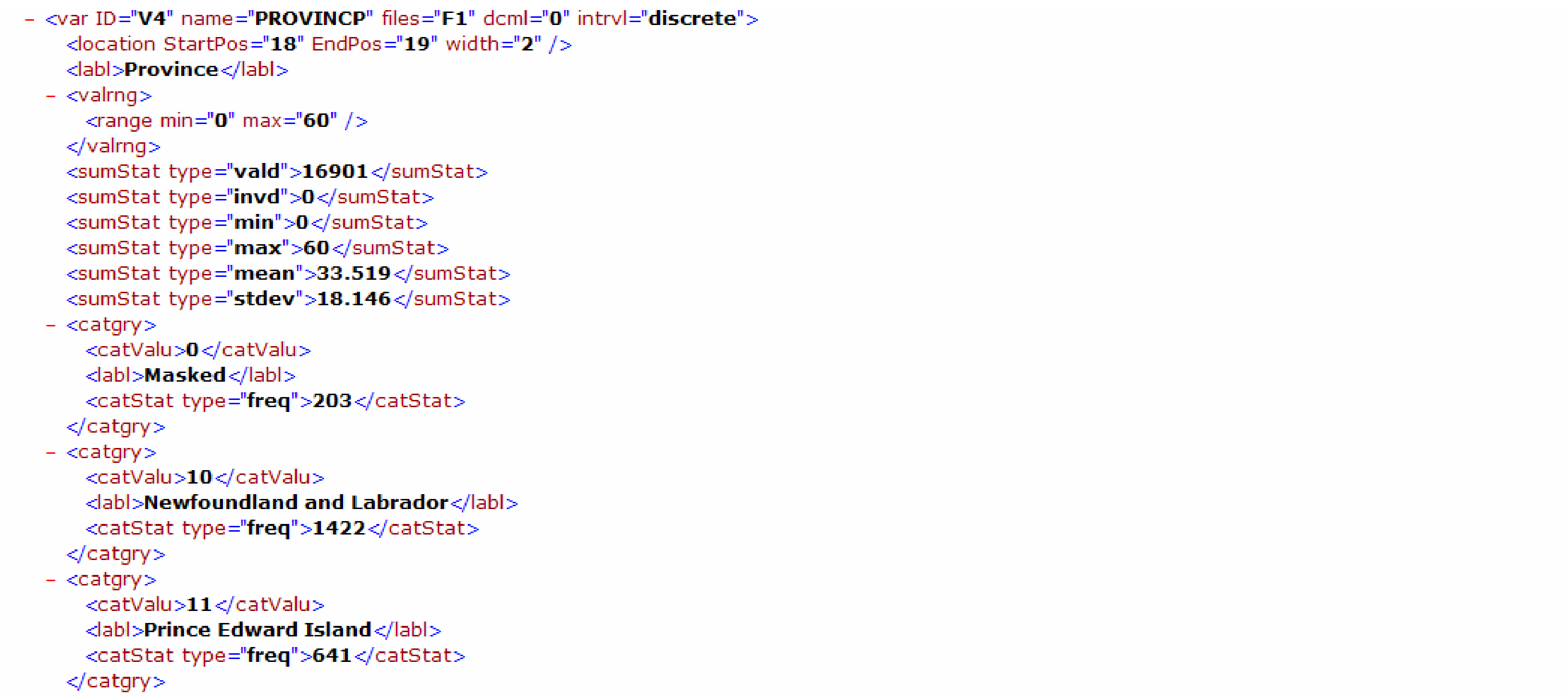
<notes>**Derived from PA_Q01**.</notes>

Example of Complete DDI compliant codebook for Section 4.0 Data Files Description (portion of the codebook ONLY)

SHS 2001

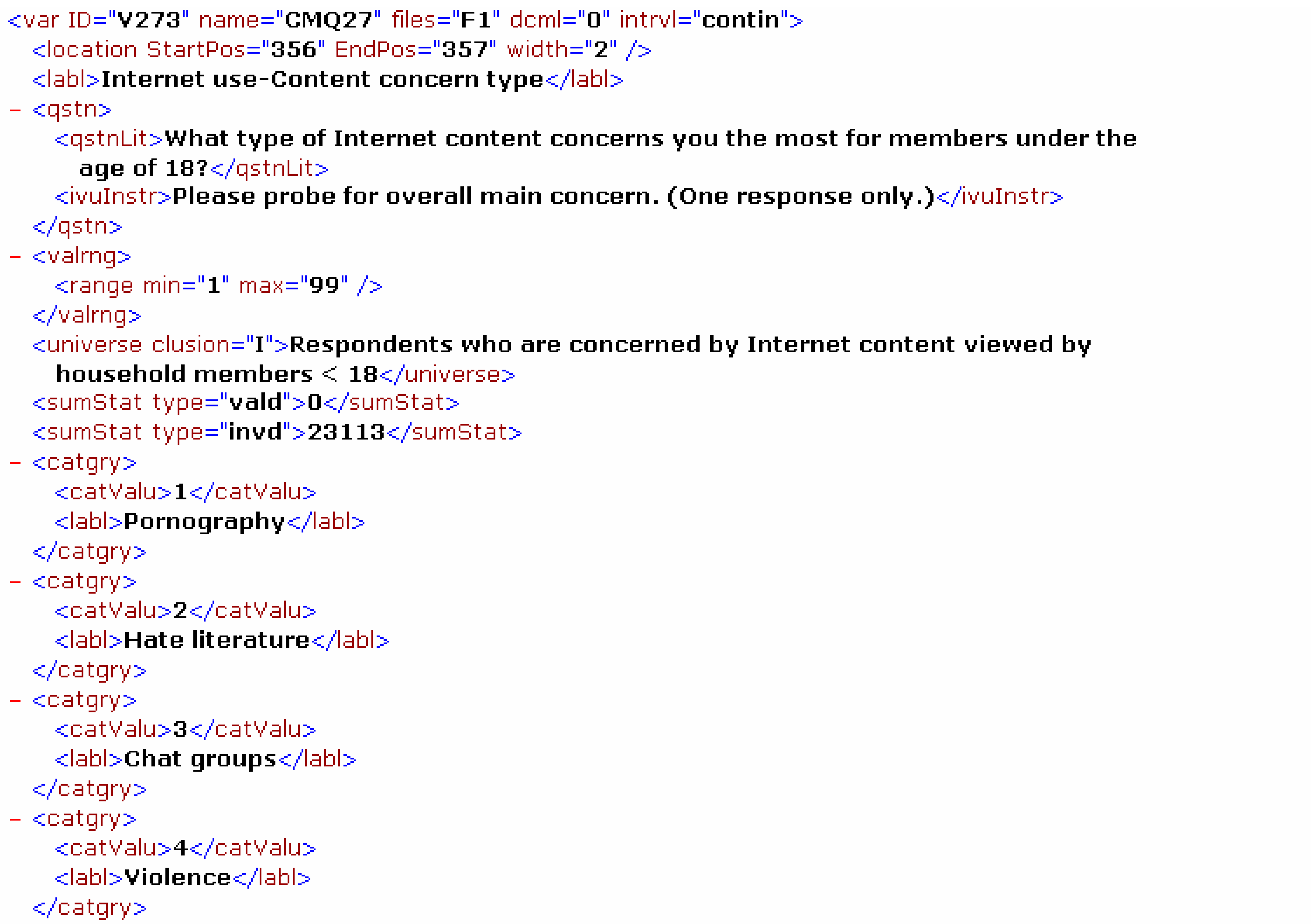


```
- <dataDscr>
  - <varGrp ID="VG1F1" type="subject" var="V1 V2 V3 V4 V5 V6">
      <labl>Location</labl>
    </varGrp>
  - <varGrp ID="VG2F1" type="subject" var="V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20 V21 V22 V23 V24 V25 V
      <labl>Dwelling</labl>
    </varGrp>
  - <varGrp ID="VG3F1" type="subject" var="V27 V28 V29 V30 V31 V32 V33 V34 V35 V36 V37 V38">
      <labl>Characteristics of Reference Person</labl>
    </varGrp>
```



```
- <var ID="V4" name="PROVINCP" files="F1" dcml="0" intrvl="discrete">
    <location StartPos="18" EndPos="19" width="2" />
    <labl>Province</labl>
  - <valrng>
      <range min="0" max="60" />
    </valrng>
    <sumStat type="vald">16901</sumStat>
    <sumStat type="invd">0</sumStat>
    <sumStat type="min">0</sumStat>
    <sumStat type="max">60</sumStat>
    <sumStat type="mean">33.519</sumStat>
    <sumStat type="stdev">18.146</sumStat>
  - <catgry>
      <catValu>0</catValu>
      <labl>Masked</labl>
      <catStat type="freq">203</catStat>
    </catgry>
  - <catgry>
      <catValu>10</catValu>
      <labl>Newfoundland and Labrador</labl>
      <catStat type="freq">1422</catStat>
    </catgry>
  - <catgry>
      <catValu>11</catValu>
      <labl>Prince Edward Island</labl>
      <catStat type="freq">641</catStat>
    </catgry>
```

HIUS 2003



```
<var ID="V273" name="CMQ27" files="F1" dcml="0" intrvl="contin">
  <location StartPos="356" EndPos="357" width="2" />
  <labl>Internet use-Content concern type</labl>
- <qstn>
    <qstnLit>What type of Internet content concerns you the most for members under the
      age of 18?</qstnLit>
    <ivuInstr>Please probe for overall main concern. (One response only.)</ivuInstr>
  </qstn>
- <valrng>
    <range min="1" max="99" />
  </valrng>
  <universe clusion="I">Respondents who are concerned by Internet content viewed by
    household members < 18</universe>
  <sumStat type="vald">0</sumStat>
  <sumStat type="invd">23113</sumStat>
- <catgry>
    <catValu>1</catValu>
    <labl>Pornography</labl>
  </catgry>
- <catgry>
    <catValu>2</catValu>
    <labl>Hate literature</labl>
  </catgry>
- <catgry>
    <catValu>3</catValu>
    <labl>Chat groups</labl>
  </catgry>
- <catgry>
    <catValu>4</catValu>
    <labl>Violence</labl>
  </catgry>
```