

SAS[®] GLOBAL FORUM 2018

USERS PROGRAM

Extracting Data from Zipped XML Files with the ZIP and XML FILENAME Options in SAS[®]

Andrew Thomson, Maura Bardos

Energy Information Administration Washington, DC

April 8 - 11 | Denver, CO

#SASGF

Extracting Data from Zipped XML Files with the ZIP and XML FILENAME Options in SAS®

Andrew Thomson, Maura Bardos

Energy Information Administration

Abstract

One of the challenges facing statistical agencies is leveraging large volumes of transactional data. A common means of transmitting this type of information is in zipped batches of XML files. This poster describes how the SAS® FILENAME ZIP access method can be used to first read in the XML files from a zipped folder, and then read in and parse the XML data using the FILENAME XML option. This poster also describes the use of the XML Mapper program to generate a schema to parse the data according to the programmer's specification. Finally, SAS macro language techniques are applied to further automate the process across all XML files in all zipped folders in a given directory.

This eposter will provide users a high-level, abbreviated walk through to read in XML files and parse out their data elements.

Reading from a zipped folder

1. Begin by setting an INFILE to point to a zipped folder and use the ZIP access method

- Here we call the filename directory "zips", pointing to the actual zipped folder

```
filename zips "C:\SASUniversityEdition\myfolders\In\batch1.zip";
```

2. Use directory functions DOPEN, DNUM, DREAD and DCLOSE to read in files from the zipped folder under a data step.

- We create a dataset called "zipped_contents" which contains the name and index number of each xml file inside "zips"

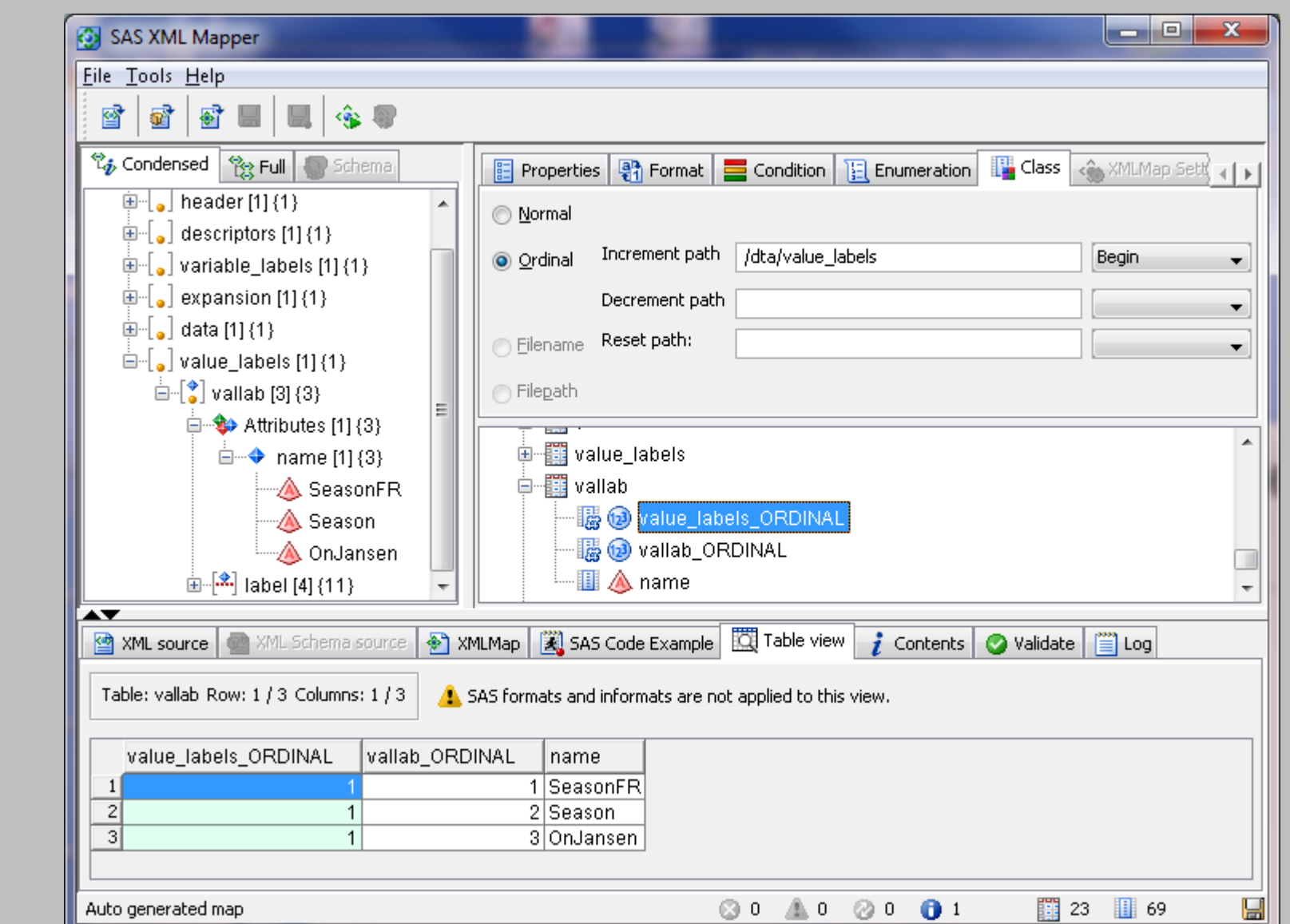
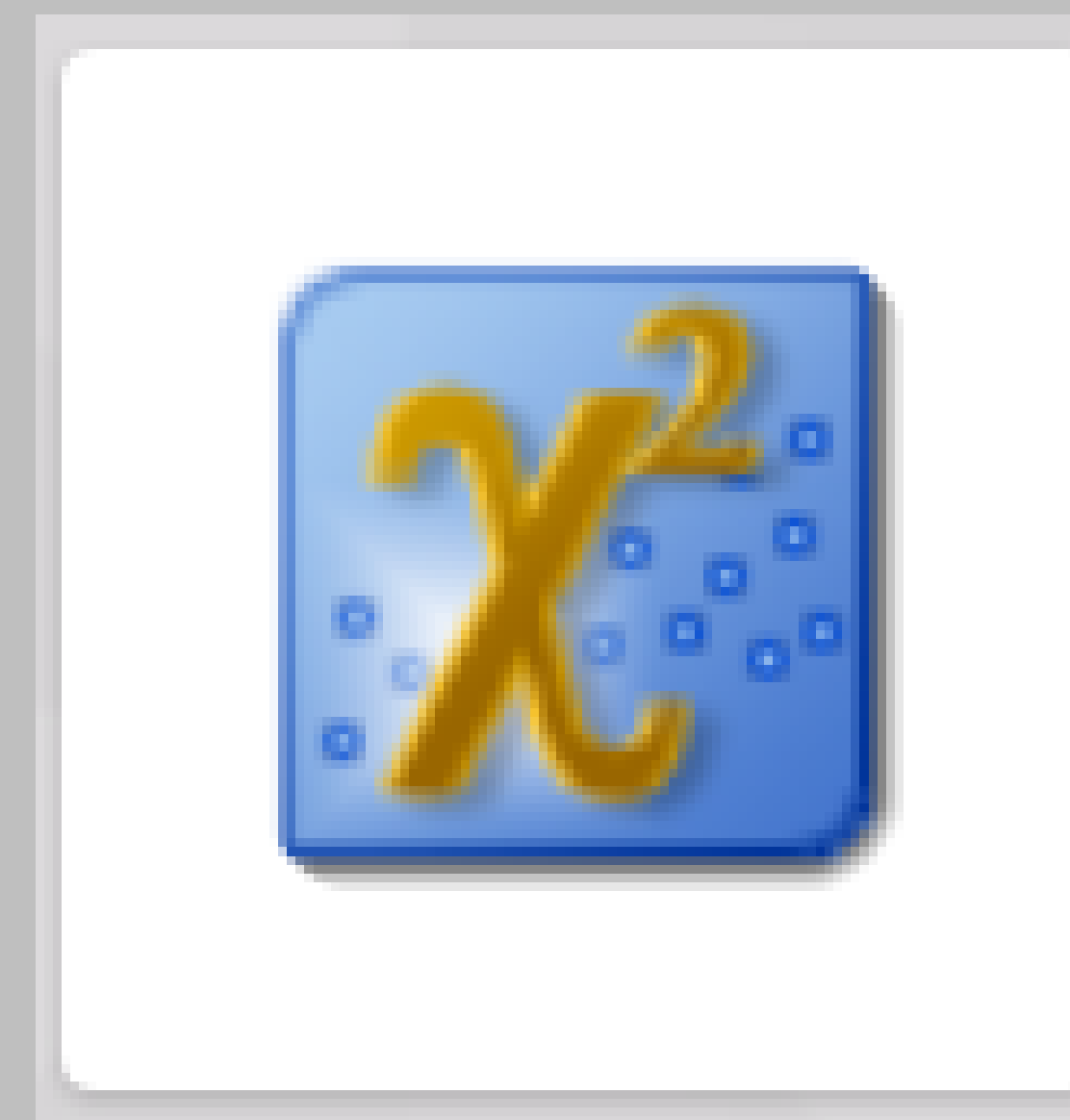
```
*get names of zip files contained within the directory;
data zips_contents;
  fid=dopen("zips");
  if fid=0 then stop;

  memcount = dnum(fid);

  do i = 1 to memcount;
    memname = dread(fid,i);
    output;
  end;
run;
```

Reading in XML files

1. Use the SAS XML Mapper to create a Map to help SAS parse out the data elements in the XML file.
 - This is a free download from SAS and generate a map to guide SAS to parse out the data from the XML files



2. Use a FILENAME to read in the map

- A sample XML file was fed into the SAS XML Mapper to create a map called "my_map.map"
- We call the filename directory "SXLEMAP" that points to the XML map file

```
filename SXLEMAP ' C:\SASUniversityEdition\myfolders\Maps\my_map.map';
```

Extracting Data from Zipped XML Files with the ZIP and XML FILENAME Options in SAS®

Andrew Thomson, Maura Bardos

Energy Information Administration

Parsing out XML data

1. Use a LIBNAME with XMLV2 method to read in the XML file and the XMLMAP= to help SAS parse out the data elements.
 - This libname called “my_xml” acts as a temporary staging area for the parsed out data from each xml file before being read to a permanent dataset

```
libname my_xml XMLV2 "%sysfunc(getoption(work))\&&xmlname&j" xmlmap=sxlemap  
access=readonly;
```

2. This dataset called “xml_test” acts as the permanent staging area where data elements from each xml file appended.
 - Note that it used the “my_xml” libname.
 - The “extracted_data” table is the name derived from the xml map.
 - From here, the parsed data in “xml_test” can be manipulated and used later in the program

```
data xml_test;  
    set my_xml.extracted_data;  
run;
```

Looping it all together

Nested DO loop:

1. Loop through 1.. i zipped folders
 - Use proc sql to create macro variables for each zipped folder

```
proc sql noprint;  
select count(*)  
into :zipNObs  
from zips_contents;  
select memname  
into :zipname1-:zipname%left(&zipNObs)  
from zips_contents;  
quit;
```

2. Loop through 1.. j XML files in each folder
 - Use proc sql to create macro variables for each XML file within the ith folder

```
proc sql noprint;  
select count(*)  
into :xmlNObs  
from contents_  
select memname  
into :xmlname1-:xmlname%left(&xmlNObs)  
from contents_  
quit;
```

REFERENCES

- Hemedinger, Chris 2015. “Using FILENAME ZIP to unzip and read data files in SAS.” The SAS Dummy. SAS Institute Inc. Cary, NC. <https://blogs.sas.com/content/sasdummys/2015/05/11/using-filename-zip-to-unzip-and-read-data-files-in-sas/>
- Cochran, Ben. “Reading and processing the Contents of a Directory.” The Bedford Group. Raleigh, NC.
- Zhao, Yi 2012. “Import and Output XML Files with SAS.” SAS Global Forum 2012 Programming: Foundations and Fundamentals, 253-2012. Merck sharp & Dohme Corp. Upper Gwenedd, PA.
- Martell, Carol. “SAS XML Mapper to the Rescue.” SESUG Proceedings. University of North Carolina Highway Safety Research Center. Chapel Hill, NC.
- Cox, Thomas 2012. “Advanced XML Processing with SAS 9.3.” SAS Global forum 2012 Programming: Beyond the Basics, 220-2012. SAS Institute Inc. Cary, NC.
- Huang, Louie and Guzman-Becerra, Norma 2009. “PROC SQL and SAS Macro: Beyond the Basics.” SAS Global Forum 2009 Posters, 200-2009. Baster Healthcare Corporation. Westlake Village, CA.
- Satchi, Thiru. “Using the Magical Keyword ‘Into’” in PROC SQL.” SUGI 27 Coders’ Corner, 71-27. Blue Cross Blue Shield of Massachusetts. Boston, MA.



SAS[®] GLOBAL FORUM 2018

April 8 - 11 | Denver, CO
Colorado Convention Center

#SASGF