**Reading and Writing NAACCR V22 XML Data Files Using SAS**

***Update 2022-09-26 for v9.0***

**Step 1. Learn about NAACCR XML.**

Before you get started, you may want to read more about the NAACCR XML standard, here:

<https://www.naaccr.org/xml-data-exchange-standard/>

**Step 2. “Git” some tools.**

SAS and XML are not best of friends, so reading and writing NAACCR V21+ files using SAS is more involved than the previous NAACCR versions that supported flat files. In order to help SAS and XML get along, you will need some tools, including SAS Macros, that were written by Fabian Depry (IMS). [***Thank you, Fabian!***] Navigate to the NAACCR-XML Java Library on GitHub (https://github.com/imsweb/naaccr-xml). The GitHub project will allow you to access a Java library for reading and writing NAACCR XML.

1. Download and “install” the latest version of the naaccr-xml-utility file from:

<https://github.com/imsweb/naaccr-xml/releases>. You only need to download the NAACCR XML Utility ZIP file, not the source code files. As of 2022-09-26, the latest version is 9.0.

1. Save the utility in any folder you like and unzip it. When you unzip the file, you will see the following files in the destination folder: 

To run the NAACCR XML Utility software, you only need to double-click on the “NAACCR XML Utility.exe.”

**Step 3. Prepare your XML data file and the user-defined dictionary file that accompanies it (if there is one).** This step “*depends*.” It *depends* on your registry software and it *depends* on the purpose for creating the file. Below are:

(a) an example using **SEER\*DMS**,

(b) comments for **CRS Plus**, and

(c) comments if you were **“just given”** a NAACCR XML data file, e.g. for a research project.

The XML data file and user-defined dictionary (likely also in XML format) come as a pair; you need the dictionary to read the data file. If the data file is a 100% standard NAACCR XML data file (Type **I**ncidence, **C**onfidential, or **A**bstract), you will only need the standard data dictionary to read your data file and will not need an additional user-defined dictionary. This is true even if your data file includes only a subset of the standard NAACCR data items. However, if your data file has any extra data items, there needs to be a second user-defined dictionary. The SEER, NPCR, and NAACCR data submissions all include non-standard fields, so all have these second user-defined dictionaries. The standard data dictionary is baked into the SAS program.

The standard NAACCR and user-defined dictionaries include a Uniform Resource Identifier (URI). The URI can be found as a root attribute of the XML dictionary. Note that the dictionary URI values might look like internet addresses, but in general they don't point to an existing web location. That's because URIs are not URLs (Uniform Resource Locators) but they often use the same convention: a path delimited by slashes, with the beginning of the path representing an organization and each remaining part of the path representing a more specific part of the resource. A given dictionary URI can point to an actual web location containing the dictionary, but that is not a requirement, and the URI of the standard NAACCR base dictionaries don't reference actual locations. Don't worry; your data are not being sent to a website.

a. This example is for registries using **SEER\*DMS** who want to use SAS on a SEER Submission Extract XML file.

In SEER\*DMS, navigate to View -> Reports and select EXT-02, SEER Submission Extract. After you have selected the appropriate parameters, click “Run Offline.”



When the extract has completed successfully, download the data file, which will be G-zipped. Also download the user-defined data dictionary, which will be in both CSV and XML format. Note that the SEER\*DMS EXT-02 Extract already has both the base NAACCR dictionary and the user-defined submission dictionary in CSV format, so you can skip Step 4 below. Step 4 may be needed for other SEER\*DMS extracts, so is included for completeness.



The SAS Macro can read XML data files in an unzipped or GZIP format, so you do not need to unzip the GZIP version that will be downloaded from SEER\*DMS.

b. For Registry Plus, the Registry Plus Call for Data Extract Program for CRS Plus users will create a NAACCR XML file for the 2022 data submission to NAACCR and NPCR CSS. All extracts created in CRS Plus V22 will be NAACCR V22 XML data files. NAACCR XML or delimited files can be created using the CRS Plus Extract Wizard to generate data files for researchers. Instructions are in the CRS Plus User’s Manual. Please contact cancerinformatics@cdc.gov if there are any questions regarding Registry Plus products.

c. If you are a researcher or other who was **“just given”** a NAACCR XML data file, you may or may not need a second user-defined dictionary. If there were fields in the data file other than standard NAACCR fields, you will need a user-defined dictionary, which the person providing the data files should be able to provide.

**Step 4. Convert the dictionaries from XML to CSV.**

Dictionaries are usually in XML format, but for technical reasons, the SAS macro expects them in CSV format. The NAACCR XML Tool that is distributed with the SAS macros has an option to load an XML dictionary and save it as CSV.

a. Open the NAACCR XML Utility.exe that you downloaded in Step 2, go to the Standard Dictionaries, and choose the NAACCR 22 base dictionary. Extract to CSV.



Click on the Standard Dictionaries icon, select the NAACCR 22 base dictionary, then Extract to CSV, navigate to where you want to save the CSV version of the base data dictionary, and Create CSV. You only need to perform these steps once, then you will have the CSV formatted version of the NAACCR 22 base dictionary for use with other NAACCR V22 XML data files.



b. This is only required if you have a user-defined dictionary, i.e. additional fields beyond what is included in the base NAACCR data dictionary. If that is the case, a dictionary should have been provided by the organization that created your XML data file or wrote the specifications for it. For example, the SEER, NPCR, and NAACCR data submission files will all include these user-defined dictionaries. Double-click on the NAACCR XML Utility.exe that you downloaded in Step 2.



Click on the Dictionary Editor icon, then the Folder Icon to select the user-defined XML dictionary you downloaded in Step 3a or 3b or were given in 3c. After the user-defined XML dictionary is loaded, click on the CSV icon, navigate to where you want to save the CSV version of the user-defined data dictionary, and Create CSV with the name you like (or default, if you like that).



**Step 5. Launch SAS.**

In SAS, open ***ReadWrite\_NAACCR\_XML\_tidy.sas*** and follow the instructions in the SAS program for modifying the code. Piece of cake, right?!

