XML Tools

Examples to Slice, Dice, Load, and Analyze XML data

Fabian Depry
Information Management Services, Inc
NAACCR 2015, Charlotte
Introduction

- Tools are important because they make complex tasks look simple
- Switching from a flat format to an XML can be a complex task
- Our Task Force developed some tools to help with that transition
Tools Overview

Flat Files
- Text Editors
- Compression Tools
- Conversion Tools
- Viewing/Processing Tools
- Abstracting Tools
- Submission Tools
- Data Management Systems
- Analysis Tools

XML
- XML Editors
- Compression Tools
## Tools Overview

<table>
<thead>
<tr>
<th>Flat Files</th>
<th>XML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text Editors</td>
<td>XML Editors</td>
</tr>
<tr>
<td>Compression Tools</td>
<td>Compression Tools</td>
</tr>
<tr>
<td>Conversion Tools</td>
<td>Conversion Tools</td>
</tr>
<tr>
<td>Viewing/Processing Tools</td>
<td>Viewing/Processing Tools</td>
</tr>
<tr>
<td>Abstracting Tools</td>
<td>Abstracting Tools</td>
</tr>
<tr>
<td>Submission Tools</td>
<td>Submission Tools</td>
</tr>
<tr>
<td>Data Management Systems</td>
<td>Data Management Systems</td>
</tr>
<tr>
<td>Analysis Tools</td>
<td>Analysis Tools</td>
</tr>
</tbody>
</table>
Tools Overview

**Flat Files**
- Text Editors
- Compression Tools
- Conversion Tools
- Viewing/Processing Tools
- Abstracting Tools
- Submission Tools
- Data Management Systems
- Analysis Tools

**XML**
- XML Editors
- Compression Tools
- Conversion Tools
- Viewing/Processing Tools
- Abstracting Tools
- Submission Tools
- Data Management Systems
- Analysis Tools
NAACCR XML Utility

- Developed by the NAACCR XML Task Force
- Freely available, open source (GitHub)
- Java standalone program, released as a single JAR
- Supports NAACCR 14 and 15 (AMCI types)
- Main features:
  - XML validation (evaluation tool)
  - Flat to XML and XML to Flat (transition tool)
  - Dictionary viewer (information tool)
GUI Overview

The following NAACCR versions are supported:
- NAACCR 15
- NAACCR 19

The Abstract, Modified, Confidential and Incidence flavors are supported for those versions.

Note that this utility is not a data conversion tool, it simply translated one format into another. That means the created file (Flat or XML) will always have the same NAACCR version (and same data) as its source.

The following compressions are supported:
- GZ2 (".gz" extension)
- XZ (".xz" extension); this compression will usually produce smaller files than GZ2 but will take longer to process
- Uncompressed (anything not ending in ".gz" or ".xz" will be treated as uncompressed)
Going from Flat to XML (step 1 of 7)

Source Flat File:

No file selected; please use the Browse button to select one.

The following NAACCR versions are supported:
NAACCR 15
NAACCR 14

The Abstract, Modified, Confidential and Incidence flavors are supported for those versions.

Note that this utility is not a data conversion tool, it simply translated one format into another.
That means the created file (Flat or XML) will always have the same NAACCR version (and same data) as its source.

The following compressions are supported:
GZip (*.gz extension)
XZ (*.xz extension; this compression will usually produce smaller files than GZip but will take longer to process
Uncompressed (anything not ending in .gz or .xz will be treated as uncompressed)
Going from Flat to XML (step 2 of 7)

Source Flat File: C:\Users\depryf\Desktop\fake-naaccr14inc-10000-rec.txt.gz

Source File format: Compressed NAACCR 14.0 Incidence  Number of lines: <not evaluated yet>  File size: 3.6 MB

Target XML File: C:\Users\depryf\Desktop\fake-naaccr14inc-10000-rec.xml.gz

Please review and/or change the following options. Once you are ready, click the process button at the bottom of the page.
Processing Options

- When reading the tumors, group them by Patient ID Number (Item #20).
  
  *If this option is checked, the tumors will be grouped together, resulting in several tumors per patient.*
  Otherwise the tumors won’t be grouped and every patient will contain exactly one tumor.
  
  *If this option is selected, the lines in the flat file belonging to the same patient are assumed to appear next to each other.*

- When grouping the tumors, report value mismatch.
  
  *If this option is checked, the items of the tumors grouped together, but having different values will be reported as warnings.*
  The few items defined as root-items (like registry ID) but having different values for different patients will also be reported.

- When reading the items, validate their value.
  
  *If this option is checked, each value will be validated against the item's data type defined in the dictionary.*

- When writing the items in the XML file, also include the NAACCR Number.
  
  *If this option is checked, the NAACCR Numbers will be written to the file in addition to the NAACCR IDs.*
  Otherwise only the NAACCR ID (which is required) is written as an attribute.
Going from Flat to XML (step 4 of 7)
Going from Flat to XML (step 5 of 7)
### Going from Flat to XML (step 6 of 7)

Source Flat File: `C:\Users\depry\Desktop\fake-naaccr14-inc-10000-rec.txt.gz`

**Source File format:** Compressed NAACCR 14.0 Incidence  
**Number of lines:** 10,000  
**File size:** 3.6 MB

Successfully created "C:\Users\depry\Desktop\fake-naaccr14-inc-10000-rec.xml.gz" (11.5 MB) in 8 seconds (analysis: < 1 second, processing: 7 seconds)

<table>
<thead>
<tr>
<th>Warnings</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line 1, item 'dateOfBirth' (#240): invalid value according to data type 'date': 28482365</td>
<td></td>
</tr>
<tr>
<td>Line 1, item 'dateOfLastContact' (#1750): invalid value according to data type 'date': 20100687</td>
<td></td>
</tr>
<tr>
<td>Line 1, item 'countyCurrent' (#1840): invalid value according to data type 'alpha': 414</td>
<td></td>
</tr>
<tr>
<td>Line 1, item 'causeOfDeath' (#1910): invalid value according to data type 'digits': ABCD</td>
<td></td>
</tr>
<tr>
<td>Line 1, item 'placeOfDeath' (#1940): invalid value according to data type 'alpha': 600</td>
<td></td>
</tr>
<tr>
<td>Line 1, item 'dateOfDeathCanada' (#1755): invalid value according to data type 'date': 15160333</td>
<td></td>
</tr>
<tr>
<td>Line 1, item 'countyAtDx' (#90): invalid value according to data type 'alpha': 127</td>
<td></td>
</tr>
</tbody>
</table>
Going from Flat to XML (step 7 of 7)

Source Flat File: C:\Users\deprf\Desktop\fake-naaccr14inc-10000-rec.txt.gz

Source File format: Compressed NAACCR 14.0 Incidence  Number of lines: 10,000  File size: 3.6 MB

Successfully created "C:\Users\deprf\Desktop\fake-naaccr14inc-10000-rec.xml.gz" (11.5 MB) in 8 seconds (analysis: < 1 second, processing: 7 seconds)

Validation warning counts (0 counts not displayed):

Value too short: 10,000
  [unusualFollowUpMethod]
Value invalid for data type: 858,100
  [causeOfDeath, comorbidComplication1, comorbidComplication10, comorbidComplication2, comorbidComplication9, NaaccrData value not consistent among tumors: 9,990
  [npiRegistryId]
Going from XML to Flat

<table>
<thead>
<tr>
<th>Source XML File:</th>
<th>\Users\dep\Desktop\fake-naaccr14inc-10000-rec.xml.gz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source File format:</td>
<td>Compressed NAACCR 14.0 Incidence</td>
</tr>
<tr>
<td>Number of lines:</td>
<td>&lt;not evaluated yet&gt;</td>
</tr>
<tr>
<td>File size:</td>
<td>11.5 MB</td>
</tr>
</tbody>
</table>

Please review and/or change the following options. Once you are ready, click the process button at the bottom of the page.

<table>
<thead>
<tr>
<th>Target Flat File:</th>
<th>\Users\dep\Desktop\fake-naaccr14inc-10000-rec.txt.gz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compression:</td>
<td>GZip</td>
</tr>
</tbody>
</table>

**Processing Options**

- [ ] When reading the items, validate their value.
  
  *If this option is checked, each value will be validated against the item's data type defined in the dictionary.*

- [x] When reading the file, ignore unknown items.
  
  *If this option is checked, unknown items will be ignored. Otherwise a warning will be reported.*

**User Dictionary:**

- [ ] Browse...

- [ ] Process Source File
XML Validation

Source XML File: C:\Users\deprf\Desktop\fake-naaccr-14-inc-10000-rec.xml.gz

Source File format: Compressed NAACCR 14.0 Incidence
Number of lines: <not evaluated yet>
File size: 11.5 MB

Please review and/or change the following options. Once you are ready, click the process button at the bottom of the page.

Processing Options

- When reading the items, validate their value.
  
  *If this option is checked, each value will be validated against the item's data type defined in the dictionary.*

- When reading the file, ignore unknown items.
  
  *If this option is checked, unknown items will be ignored. Otherwise a warning will be reported.*

User Dictionary:

Process Source File
Viewing the Dictionaries

NAACCR 15 base dictionary

<?xml version="1.0"?>

<NaaccrDictionary dictionaryUri="http://naaccr.org/naaccrxml/naaccr-dictionary-150.xml"
naaccrVersion="150"
description="NAACCR 15 base dictionary"
xmns="http://naaccr.org/naaccrxml">
<ItemDefs>
  <ItemDef naaccrId="recordType"
    naaccrNum="10"
    naaccrName="Record Type"
    startColumn="1"
    length="1"
    recordTypes="A,M,C,I"
    parentXmlElement="NaaccrData"
    regexValidation="^[ICAM]$"/>
  <ItemDef naaccrId="registryType"
    naaccrNum="30"
</ItemDefs>
<table>
<thead>
<tr>
<th></th>
<th>Source</th>
<th>Record Type</th>
<th>Num Records</th>
<th>Compression</th>
<th>Time to XML (h:m:s)</th>
<th>Records Converted/sec</th>
<th>Original Size</th>
<th>Converted Size</th>
<th>Conversion Factor (Size)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Florida - full extract test file (data on network)</td>
<td>A</td>
<td>2.8M</td>
<td>None</td>
<td>3:32:00</td>
<td>220 (network)</td>
<td>62.4GB</td>
<td>25.2GB</td>
<td>0.4</td>
</tr>
<tr>
<td>2</td>
<td>Florida - full extract test file (data local)</td>
<td>A</td>
<td>2.8M</td>
<td>None</td>
<td>0:37:00</td>
<td>1261 (local)</td>
<td>62.4GB</td>
<td>25.2GB</td>
<td>0.4</td>
</tr>
<tr>
<td>3</td>
<td>Los Angeles (SEER 35) - 11/2014 sub file</td>
<td>I</td>
<td>1.0M</td>
<td>None</td>
<td>0:03:15</td>
<td>5128</td>
<td>3.1GB</td>
<td>6.6GB</td>
<td>2.1</td>
</tr>
<tr>
<td>4</td>
<td>Los Angeles (SEER 35) - 11/2014 sub</td>
<td>I</td>
<td>1.0M</td>
<td>GZ</td>
<td>0:03:05</td>
<td>5411</td>
<td>103MB</td>
<td>188MB</td>
<td>1.8</td>
</tr>
<tr>
<td>5</td>
<td>Ontario Cancer Registry - 2014</td>
<td>I</td>
<td>2.2M</td>
<td>GZ</td>
<td>0:09:00</td>
<td>4074</td>
<td>224MB</td>
<td>405MB</td>
<td>1.8</td>
</tr>
<tr>
<td>6</td>
<td>Greater Calif (SEER 41) - 2/2015 sub file</td>
<td>I</td>
<td>2.2M</td>
<td>None</td>
<td>0:05:17</td>
<td>6944</td>
<td>6.7GB</td>
<td>11.4GB</td>
<td>1.7</td>
</tr>
<tr>
<td>7</td>
<td>Ontario Cancer Registry - eMaRC Abstract</td>
<td>A</td>
<td>1567</td>
<td>None</td>
<td>0:03:50</td>
<td>7 (network)</td>
<td>35K</td>
<td>3K</td>
<td>0.1</td>
</tr>
<tr>
<td>8</td>
<td>Ontario Cancer Registry - CSA Abstracts</td>
<td>A</td>
<td>2033</td>
<td>None</td>
<td>0:05:45</td>
<td>6 (network)</td>
<td>45K</td>
<td>4K</td>
<td>0.1</td>
</tr>
<tr>
<td>9</td>
<td>Kentucky Cancer Registry - 11/2014 sub file</td>
<td>I</td>
<td>465K</td>
<td>None</td>
<td>0:01:18</td>
<td>5962</td>
<td>1.5GB</td>
<td>2.1GB</td>
<td>1.4</td>
</tr>
<tr>
<td>10</td>
<td>Kentucky Cancer Registry - 11/2014 pre sub file</td>
<td>I</td>
<td>466K</td>
<td>None</td>
<td>0:02:26</td>
<td>3196</td>
<td>2.5GB</td>
<td>5.4GB</td>
<td>2.1</td>
</tr>
<tr>
<td>11</td>
<td>Kentucky Cancer Registry - Large case file</td>
<td>A</td>
<td>566K</td>
<td>None</td>
<td>0:05:40</td>
<td>2664</td>
<td>12.6GB</td>
<td>6.0GB</td>
<td>0.5</td>
</tr>
<tr>
<td>12</td>
<td>SEER Submission data - .GZ compressed</td>
<td>I</td>
<td>2.2M</td>
<td>GZ</td>
<td>0:10:30</td>
<td>3492</td>
<td>224MB</td>
<td>404MB</td>
<td>1.8</td>
</tr>
<tr>
<td>13</td>
<td>SEER Submission data - .GZ compressed</td>
<td>I</td>
<td>468K</td>
<td>GZ</td>
<td>0:03:03</td>
<td>2557</td>
<td>76.6MB</td>
<td>188MB</td>
<td>2.5</td>
</tr>
<tr>
<td>14</td>
<td>SEER Submission data - .GZ compressed</td>
<td>I</td>
<td>201K</td>
<td>GZ</td>
<td>0:00:49</td>
<td>4102</td>
<td>18.4MB</td>
<td>33.5MB</td>
<td>1.8</td>
</tr>
<tr>
<td>15</td>
<td>SEER Submission data - .XZ compressed</td>
<td>I</td>
<td>201K</td>
<td>XZ</td>
<td>0:01:29</td>
<td>2258</td>
<td>18.4MB</td>
<td>23.9MB</td>
<td>1.2</td>
</tr>
<tr>
<td>16</td>
<td>Synthetic Test - .GZ compressed</td>
<td>A</td>
<td>2.5M</td>
<td>GZ</td>
<td>0:26:00</td>
<td>1603</td>
<td>56GB</td>
<td>61GB</td>
<td>1.1</td>
</tr>
<tr>
<td>17</td>
<td>Synthetic Test - .GZ compressed</td>
<td>A</td>
<td>2.5M</td>
<td>GZ</td>
<td>0:26:00</td>
<td>1603</td>
<td>1.3GB</td>
<td>1.7GB</td>
<td>1.3</td>
</tr>
<tr>
<td>18</td>
<td>Synthetic Test - with Grouped (nested) ON</td>
<td>A</td>
<td>25K</td>
<td>None</td>
<td>n/a</td>
<td>n/a</td>
<td>544MB</td>
<td>622MB</td>
<td>1.1</td>
</tr>
<tr>
<td>19</td>
<td>Synthetic Test - with Grouped (nested) OFF</td>
<td>A</td>
<td>25K</td>
<td>None</td>
<td>n/a</td>
<td>n/a</td>
<td>544MB</td>
<td>638MB</td>
<td>1.1</td>
</tr>
</tbody>
</table>
Other tools

- Our task force did other proof of concept with
  - SAS
  - Python
  - XML persistence (database mapping)

- But other tools will need to be written by the community
The XML format will work, if it’s embraced by NAACCR and the community.

More work needs to be done, and not only in terms of design and implementation, but also decisions about the logical level of the items.

The Task Force will keep improving the tools to make the transition as smooth as possible!
Conclusion

Thank you!

Check out our work and/or download the tool:

http://naacrcrxml.org/