

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

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

Tools Overview

Flat Files

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



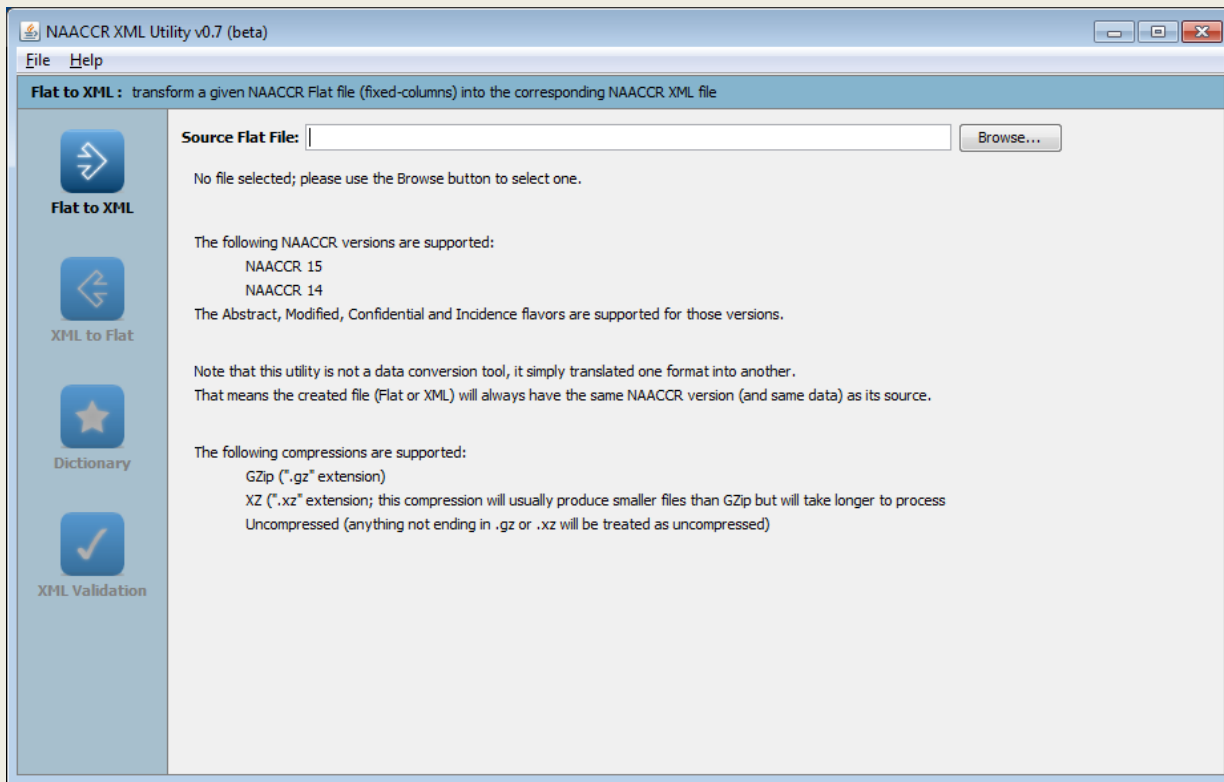
XML

- 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



Going from Flat to XML (step 1 of 7)

Source Flat File:

Browse...

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:

Source File format: Compressed NAACCR 14.0 Incidence

Number of lines: <not evaluated yet>

File size: 3.6 MB

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

Target XML File:

Compression:

Going from Flat to XML (step 3 of 7)

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)

User Dictionary:

Going from Flat to XML (step 5 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

Analyzing file (this can take a while, especially when reading network resources)...

Source Flat File: C:\Users\depryf\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

Processing file...

Going from Flat to XML (step 6 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:** 10,000 **File size:** 3.6 MB

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

Warnings	Summary
Line 1, item 'dateOfBirth' (#240): invalid value according to data type 'date': 28482365	
Line 1, item 'dateOfLastContact' (#1750): invalid value according to data type 'date': 20100687	
Line 1, item 'countyCurrent' (#1840): invalid value according to data type 'alpha': 414	
Line 1, item 'causeOfDeath' (#1910): invalid value according to data type 'digits': ABCD	
Line 1, item 'placeOfDeath' (#1940): invalid value according to data type 'alpha': 600	
Line 1, item 'dateOfDeathCanada' (#1755): invalid value according to data type 'date': 15160333	
Line 1, item 'countyAtDx' (#90): invalid value according to data type 'alpha': 127	

Going from Flat to XML (step 7 of 7)

Source Flat File:

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

Successfully created "C:\Users\depryf\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, comorbidCompli

NaaccrData value not consistent among tumors: 9,990

[npiRegistryId]

Going from XML to Flat

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

Browse...

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.

Target Flat File: C:\Users\depryf\Desktop\fake-naaccr 14inc-10000-rec.txt.gz

Browse...

Compression: GZip

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:

Browse...

Process Source File

XML Validation

Source XML File:

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:

Viewing the Dictionaries

NAACCR 15 base dictionary <http://naaccr.org/naaccrxml/naaccr-dictionary-150.xml>

```
<?xml version="1.0"?>
<NaaccrDictionary dictionaryUri="http://naaccr.org/naaccrxml/naaccr-dictionary-150.xml"
  naaccrVersion="150"
  description="NAACCR 15 base dictionary"
  xmlns="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"
```

NAACCR 15 base dictionary <http://naaccr.org/naaccrxml/naaccr-dictionary-150.xml>

- NAACCR 14 base dictionary
- <NAACCR 14 default user dictionary
- NAACCR 15 base dictionary
- <NAACCR 15 default user dictionary

```
  dictionaryUri="http://naaccr.org/naaccrxml/naaccr-dictionary-150.xml"
  naaccrVersion="150"
  description="NAACCR 15 base dictionary"
  xmlns="http://naaccr.org/naaccrxml">
  <ItemDefs>
```

Evaluating Speed and Size

	Source	Record Type	Num Records	Compression	Time to XML (h:m:s)	Records Converted/sec	Original Size	Converted Size	Conversion Factor (Size)
1	Florida - full extract test file (data on network)	A	2.8M	None	3:32:00	220 (network)	62.4GB	25.2GB	0.4
2	Florida - full extract test file (data local)	A	2.8M	None	0:37:00	1261 (local)	62.4GB	25.2GB	0.4
3	Los Angeles (SEER 35) - 11/2014 sub file	I	1.0M	None	0:03:15	5128	3.1GB	6.6GB	2.1
4	Los Angeles (SEER 35) - 11/2014 sub	I	1.0M	GZ	0:03:05	5411	103MB	188MB	1.8
5	Ontario Cancer Registry - 2014	I	2.2M	GZ	0:09:00	4074	224MB	405MB	1.8
6	Greater Calif (SEER 41) - 2/2015 sub file	I	2.2M	None	0:05:17	6944	6.7GB	11.4GB	1.7
7	Ontario Cancer Registry - eMaRC Abstract	A	1567	None	0:03:50	7 (network)	35K	3K	0.1
8	Ontario Cancer Registry - CSA Abstracts	A	2033	None	0:05:45	6 (network)	45K	4K	0.1
9	Kentucky Cancer Registry - 11/2014 sub file	I	465K	None	0:01:18	5962	1.5GB	2.1GB	1.4
10	Kentucky Cancer Registry - 11/2014 pre sub file	C	466K	None	0:02:26	3196	2.5GB	5.4GB	2.1
11	Kentucky Cancer Registry - Large case file	A	566K	None	0:05:40	2664	12.6GB	6.0GB	0.5
12	SEER Submission data - .GZ compressed	I	2.2M	GZ	0:10:30	3492	224MB	404MB	1.8
13	SEER Submission data - .GZ compressed	I	468K	GZ	0:03:03	2557	76.6MB	188MB	2.5
14	SEER Submission data - .GZ compressed	I	201K	GZ	0:00:49	4102	18.4MB	33.5MB	1.8
15	SEER Submission data - .XZ compressed	I	201K	XZ	0:01:29	2258	18.4MB	23.9MB	1.2
16	Synthetic Test	A	2.5M	None	0:26:00	1603	56GB	61GB	1.1
17	Synthetic Test - .GZ compressed	A	2.5M	GZ	0:26:00	1603	1.3GB	1.7GB	1.3
18	Synthetic Test - with Grouped (nested) ON	A	25K	None	n/a	n/a	544MB	622MB	1.1
19	Synthetic Test - with Grouped (nested) OFF	A	25K	None	n/a	n/a	544MB	638MB	1.1



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

Conclusion

- 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://naaccrxml.org/>