Part 1 – Question 6

Here are the steps of canonicalization according to Data Curation coursework:

1. Convert to a single character encoding and normalize line ends
2. Remove all comments, tabs, non-significant spaces, etc.
3. Propagate all attributes defaults indicated in the schema to the elements themselves
4. Put attribute/value pairs on elements in alpha order
5. Expand all character references
6. Remove any internal schema or declarations
7. Now test to see if character sequences are identical
8. Describe your process for canonicalization (i.e., decisions, actions, representation selection, attribute issues, provenance decisions). Report the checksum values after canonicalization.

Here are the steps that I followed:

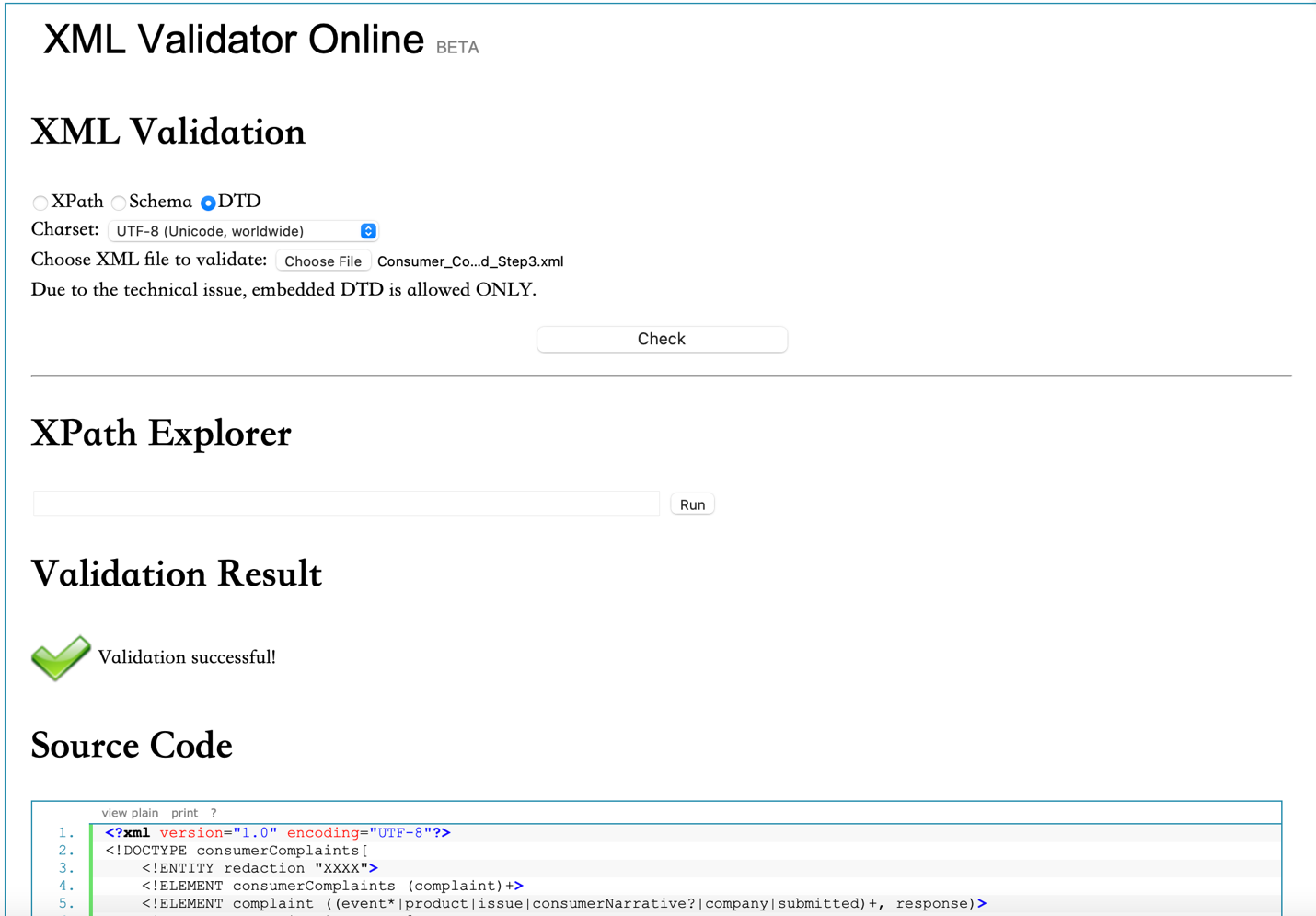
* + Both File A and File B used UTF-8 encoding, no change is needed
  + Remove submitted elements and add attribute “submissionType” in complaint elements
  + Trimming all the whitespace inside the bracket. From <example > to <example>
  + Self-closing tags are replaced by open and close tags. From <example/> to <example></example>
  + Reordering the attributes in elements to alphabetical order in response and event elements.
  + Newline and spacing corrected in “CustomerNarrative”. From paragraph into one line.
  + Change element response attribute timely from yes/no to Y/N
  + Comment deleted for File B

After completed these steps above the MD5 checksum didn’t match, so compare and match the files by the following steps:

* Adding entity and rename “&redaction;” to “XXXX” in File A
* Remove empty tag <submitted></submitted> in File B
* Add timely attribute, timely attribute is missing in File B the 7th and 8th complaints

Please see the attached files:

Consumer\_Complaints\_FileA\_Canonicalized\_Step3.xml



MD5 Checksum of File A canonicalized: 7BF00E291AC30E4D623993FB46EFFBBA

Consumer\_Complaints\_FileB\_Canonicalized\_Step3.xml

Graphical user interface, text, application, email

Description automatically generated

MD5 Checksum of File B canonicalized: 7BF00E291AC30E4D623993FB46EFFBBA

1. How does the way data is represented impact reproducibility?

Reproducibility is defined as the ability to reproduce data / results by ensuring the validity and reliability. File A and file B also have incomplete or missing the DTD. Compare with File A, File B also have missing information. Canonicalized steps can help us to standardize the procedures. They can provide the same data representation with same data despite different format. Having same data representation would aid the reader and user more efficient to identify and analyze the data. Furthermore, it also helps sharing the data to others more easily.

1. How may your canonicalization support the overarching goals of data curation (revisit objectives and activities of Week 1)?

* Organization: DTD and defined standard help organize the data
* Preservation: Clear and well-defined Metadata documentation DTD would let the data easy to understand and usable in the future
* Identification: xml file supports the ability to identify, authenticate, and validate data
* Sharing: standardization of file would aid others readable and enhance sharing
* Reproducibility: by following the canonicalized steps, we can reproduce the result with different format of input easily.

1. Which additional curation activities would you recommend to enhance the data set for future discovery and use?

* Security: version control and modification need to be specified
* Compliance: make sure follows state and federal privacy regulations
* Storage: cloud storage can be used for reliability and accessibility